

Estimating Air Pollution and its Long and Short Run
Causality with Chronic Diseases
Using Vector Error Correction Model (VECM)

Ahmad R. Alsaber

Department of Mathematics and Statistics

University of Strathclyde
Glasgow, G1 1XH, UK

This thesis is submitted to the University of Strathclyde for the
degree of Doctor of Philosophy in the Faculty of Science

August 24, 2022

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Acknowledgements

First of all, I would like to express my sincere thanks and gratitude to Allah, who gave me the patience, strength, and ability to finish this work.

I am extremely thankful to my supervisor, Dr. Jiazhu Pan. He has helped, supported, and encouraged me in everything during the course of my studies.

I would also like to thank Dr. Adeeba Al-Herz for her support in providing me the data from the Kuwait Registry for Rheumatic Diseases (KRRD).

My thanks also go out to the Environmental Public Authority of Kuwait (K-EPA) and especially Mrs. Noura Al-Bannay for providing me with the environmental information regarding air quality data from the Environmental Monitoring Information System of Kuwait (eMISK) that has been essential to this research.

I am tremendously thankful to Dr. Adeeba Al-Harban from Kuwait University's Department of Earth and Environmental Sciences for her academic support; and to Dr. Rajesh Rajan for his support in providing medical information related to my research requirements. I am also very grateful to, Drs. Parul Setiya, Hamad Alaslawi, Coffie Emmanuel, Ralph Palliam, and Adnan Al-Ali; Eng. Ebraheem AL-Foraih; and Ms. Maryam Zarei, who all helped me in numerous ways during various stages of my Ph.D. study. And, of course, thanks to Mr. Jeremiah K Garrett, for helping me enormously, especially with language editing and proofreading.

Finally, I would like to thank my mum, dad, and my supportive wife, Nouf, with my three wonderful children, Jood, Mohammad and Zainah, who provide unending inspiration. I salute you all for your selfless love, care, pain, and sacrifice that have shaped my life.

Abstract

Air pollution has been linked to a number of health impacts and has been studied in a variety of contexts using a variety of studies and methodologies. This thesis is made up of a collection of papers that cover a wide range of research subjects and illustrate different study analysis and design methodologies. Multiple imputation (MI) techniques were used to deal with the missing data, where missForest had the lowest imputation error among the other imputation approaches. Time series modelling was used to predict Rheumatoid Arthritis (RA) disease activity score (DAS28) using the information of air pollution. This thesis examined the linkage among SO_2 , NO_2 , O_3 and disease activity scores for patients with RA in Kuwait. The association was investigated using the Granger causality test (using the VECM approach and other time series approaches) (in analysis of static causality) and the Impulse Response Functions (IRFs) analysis (in analysis of dynamic causality). A comprehensive conceptual framework was used in the study, which included a cointegration test, unit root test, and panel VECM. Long-run causation and asymptotic convergence among the variables were determined using the panel VECM. The empirical outcomes show that NO_2 and O_3 are statistically significant in cases when DAS28 is the dependent variable, in most of the study locations (ASA, FAH, MAN and JAH). The results demonstrate that the lagged error correction term (ECT) coefficients in DAS28 and air pollution emissions are statistically significant. Overall, the main conclusion found in this thesis and according to the cointegration test, the results show that there exists a long run relationship between the emissions of air pollution and the change of DAS28 among RA patients.

Contents

Acknowledgement	ii
Abstract	iii
List of Figures	x
List of Tables	xix
List of Abbreviations	xxvi
1 Introduction	1
1.1 Background	1
1.2 Ambient Air Pollution	5
1.3 Fundamentals of Ambient Air Pollution	6
1.3.1 Sulphur Dioxide (SO_2)	7
1.3.2 Particulate Matter (PM_{10} or $PM_{2.5}$)	8
1.3.3 Ozone (O_3)	9
1.3.4 Nitrogen Dioxide (NO_2)	10
1.3.5 Air Quality Index (AQI)	11
1.4 Studying the Relationship between Health and Air Pollution	14
1.5 Aim of the Study	15
1.5.1 Significance of Study	17
1.5.2 The Study Region and Data	18
1.5.3 Research Objective	18

Contents

1.6	Structure of the Thesis	19
2	Exploring the Characteristics of Air Pollution	22
2.1	Introduction	22
2.2	Air Quality Studies	23
2.3	Data and Methods	25
2.3.1	Description of the Study Area	25
2.3.2	Air Quality Data Collection	25
2.3.3	Statistical Analysis	26
2.4	Air Pollution Results in Kuwait	28
2.4.1	Description of Exposure Data	39
2.4.2	Conclusion of Air Quality Assessment in Kuwait	41
3	The Association between the Rheumatoid Arthritis Disease Activity Score and the Ambient Air Pollution	43
3.1	Introduction	44
3.2	Materials and Methods	46
3.2.1	Data on RA from the Kuwait Registry for Rheumatic Diseases (KRRD)	46
3.2.2	Calculating RA Indices	46
3.2.3	Ambient Ambient air Pollutants' Data (Environmental Public Authority of Kuwait—K-EPA)	47
3.2.4	Air Pollution Data Processing and Treatment	48
3.2.5	Matching Procedure between Patients and AQI	49
3.2.6	Statistical Analysis	48
3.3	Results of RA Patient Characteristics and Air Pollution Relationship	50
3.4	Discussion	58
3.5	Conclusions	62
4	Dealing with Environmental and Clinical Missing Data	63

Contents

4.1	Study 1: Dealing with Environmental Missing Data - Application on K-EPA Data	64
4.1.1	Missing Data Imputation for the K-EPA Dataset	65
4.1.2	The Review of Missing Imputation	66
4.1.3	Missing Data Mechanisms	67
4.1.4	Ignoring the Missing Data Mechanism	70
4.1.5	Multiple Imputation (MI)	71
4.1.6	Multiple Imputation Using Random Forest Method	73
4.1.7	Process of Multiple Imputations (MI) Using Rubin's Rules	76
4.1.8	Data Sets from Kuwait EPA	78
4.1.9	Missing Imputation Evaluation Criteria	79
4.1.10	R Packages Used for Imputation Process	79
4.1.11	Statistical Results	80
4.1.12	Air Quality Missing Data Patterns	86
4.1.13	Study 1 - Discussion and Conclusion	89
4.2	Study 2: Dealing with Clinical Missing Data - Application on KRRD	93
4.2.1	Missing Imputation - Rubin's Approach	94
4.2.2	Categorisation of Missing Values	94
4.2.3	Methods Used for Imputing Missing Values	95
4.2.4	Data Source—Kuwait Registry for Rheumatic Diseases (KRRD)	96
4.2.5	Calculating RA Indices	97
4.2.6	Multiple Imputation (MI) Process Using Rubin's Rules	98
4.2.7	Number of Needed Imputations	99
4.2.8	Multiple Imputation Using RF_m Method	99
4.2.9	Evaluation Criteria	100
4.2.10	Study 2 Results	101
4.2.11	Predicting the Influence of RA Factors on DAS28 Using the Original Data Set	102

Contents

4.2.12	Predicting the Influence of RA Factors on DAS28 from the PMM-Imputed Data Sets	104
4.2.13	Predicting the Influence of RA Factors on DAS28 from the Imputed Data Sets Using kNN	105
4.2.14	Predicting the Influence of RA Factors on DAS28 from the RF_m -Imputed Data Sets	106
4.2.15	Predicting the Influence of RA Factors on DAS28 from the missForest-Imputed Data Sets	106
4.2.16	Study 2 - Discussion and Conclusion	109
5	Time Series Statistical Methods Review	112
5.1	Introduction	112
5.2	Fundamental Concepts	113
5.2.1	Time Series Modelling	113
5.2.2	Time Series Definition	114
5.2.3	Time Series and Stochastic Process	115
5.2.4	Correlation and Partial Correlation Matrix Functions	116
5.3	Stationary Time Series	117
5.4	Non-Stationary Time Series	119
5.5	Testing for Difference-Stationarity	120
5.6	Unit-Root Test	121
5.6.1	Dickey-Fuller Tests (DF Test and ADF Test)	122
5.6.2	Phillips-Perron Test (PP Test)	123
5.6.3	Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test	125
5.7	Autoregressive Integrated Moving Average (ARIMA) Process	126
5.8	GARCH Model	127
5.9	Vector Autoregression (VAR)	128
5.9.1	Forecasting using VAR Model	130
5.9.2	Lag Length Selection Using Information Criteria	134
5.10	Johansen and Juselius Cointegration Test	135

Contents

5.11	Granger Causality Test	138
5.12	Cointegration: Empirical Background	140
5.13	Cointegration: Model and Notation	142
5.14	Cointegration Analysis	143
5.15	VECM Theoretical Notions and the Model	145
5.16	VAR and VEC Models	147
5.17	Cointegration and VECM	150
5.18	Specification of Deterministic Terms	151
5.19	Impulse Response Functions and Variance Decompositions	153
6	The Long/Short Run Relationship Between RA and AQI Using VECM	154
6.1	Introduction	154
6.2	Background of the Study	155
6.3	Multivariate Time Series and Air Pollution	156
6.4	Importance of the Study	159
6.5	Aim and Objective	161
6.6	Procedures and Methodology	162
6.6.1	Selected Variables	162
6.6.2	Rheumatoid Arthritis (RA) Patients' Data	162
6.6.3	EPA Data and Materials	164
6.6.4	Ambient Air Quality in Kuwait	166
6.7	Descriptive Analysis and Correlation	167
6.8	Normality and Transformation Approach	171
6.8.1	Cullen and Frey Graph	172
6.9	DAS28 OLS Models	185
6.10	Observations' Time Line	188
6.11	Stationarity Test	192
6.11.1	Unit Root Test	192
6.11.2	Augmented Dicky-Fuller (ADF) Test	193

Contents

6.11.3	Augmented Dickey-Fuller Generalised Least Squares (ADF-GLS) Unit Root Test Results	194
6.11.4	KPSS Root Test Results	194
6.11.5	Phillips-Perron (PP) Test Results	196
6.12	Lag Selection Criteria	197
6.13	Johansen Cointegration Test	198
6.14	Causality Test	201
6.15	VAR Modelling Results	202
6.16	Vector Error Correction Model (VECM) Analysis	207
6.17	The Results of the Impulse Response Analysis	223
6.18	The Stability of VECM	224
6.19	Discussion and Conclusion	230
7	Multivariate Time Series: The Impact of Air Pollution on COVID-19 Daily Cases in Kuwait using the VECM Approach	234
7.1	The Relationship Between Air Pollution and COVID-19 Hospitalisation	235
7.2	Literature Review	237
7.2.1	Relationship Between Air Pollution and Human Health	237
7.2.2	Impact of Air Pollution as a Risk Factor to COVID-19 patients	237
7.2.3	The Relationship between Atmospheric Variables and Numbers of COVID-19 Cases	238
7.2.4	Time Series Analysis to Predict COVID-19 Cases	239
7.2.5	Data and Variables	239
7.2.6	Air Quality Index (AQI)	239
7.3	Results and Discussion	240
7.3.1	The Descriptive Statistics	240
7.3.2	Correlation Analysis	241
7.3.3	Results of Unit Root and Granger Causality test	243
7.3.4	Determination of Optimal VAR	248
7.4	Conclusion and Recommendations	255

Contents

8 Discussion and Conclusion	258
8.1 Overview	258
8.2 The Characteristics of Air Pollution in Kuwait	259
8.3 The Relation Between RA and Ambient Air Pollution	260
8.4 Dealing with Missing Data	261
8.4.1 Dealing with Missing Data - Air Pollution Data	261
8.4.2 Dealing with Missing Data - KRRD Data	263
8.5 Time Series Modelling to Predict COVID-19 Cases using the Information of Air Pollution	264
8.6 Time Series Modelling to Predict RA Disease Activity Score (DAS28) using the Information of Air Pollution	265
8.7 Conclusion of the Study	267
8.7.1 Limitation of the Study	267
8.7.2 Recommendation and Suggestions	269
A Air Pollution Comparisons	271
A.1 Air Pollutants' Comparison between Industrial & Residential Stations . .	271
B Missing Imputation Results	273
B.1 Figures	273
C Normality Assessment	281
Bibliography	294

List of Figures

1.1	Health impact assessment following the impact pathway chain.	6
1.2	Published works from this thesis. Thesis map details can be found in this hyperlink: https://viewer.edrawsoft.com/public/s/ae2d0447855778	21
2.1	Location map of the selected monitoring stations—modified after K-EPA eMISK 2020.	26
2.2	K-EPA mobile lab and fixed stations used for air pollution monitoring. . .	27
2.3	Time series of the studied pollutants from 2012 to 2017— EPA Kuwait. . .	35
2.4	Temporal variation of the studied pollutants according to the station site from 2012 to 2017 for NO , NO_x , NO_2 and O_3 —EPA Kuwait. The shaded areas are the 95% confidence intervals for the mean. Plots created using OpenAir in R.	36
2.5	Temporal variation of the Studied Pollutants according to the station site from 2012 to 2017 for CO , PM_{10} and $NMHC$ —EPA Kuwait. The shaded areas are the 95% confidence intervals for the mean. Plots created using OpenAir in R.	37
2.6	Temporal variation of the Studied Pollutants according to the station site from 2012 to 2017 for C_6H_6 and SO_2 —EPA Kuwait. The shaded areas are the 95% confidence intervals for the mean. Plots created using OpenAir in R.	38

List of Figures

2.7 Box plot of the monthly pollutant concentration after suitable transformation from January 2012 to December 2017. The upper whisker extends to the highest value within 1.5 IQR from the top of the rectangle, while the lower whisker extends to the lowest value within 1.5 IQR from the bottom of the rectangle. Values beyond the end of the whiskers are considered outliers and are shown as dots. 40

2.8 Air pollutant concentration according to the wind direction and wind speed from 2012 to 2017. 42

3.1 Example of eight-hour average concentrations of O_3 48

3.2 Matching procedures to combine Air Quality Index (AQI) information with Kuwait Registry for Rheumatic Diseases (KRRD) patient profile records using date and governorate address information). K-EAPL, Environmental Public Authority of Kuwait; RA, rheumatoid arthritis. 48

3.3 Air quality index (AQI) ambients for six governorate monitoring stations in Kuwait from 2013 to 2017. 54

4.1 The steps of implementing multiple imputations for PM_{10} , SO_2 , O_3 , CO , and NO_2 during 2012 to 2017 with 20 imputed datasets ($m=20$) according to site location, in the State of Kuwait. 77

4.2 Time series of air quality monitoring for SO_2 and NO_2 from 2012 to 2017, with missing values from five different locations (stations) in the State of Kuwait. 84

4.3 Time series of air quality monitoring for O_3 and CO from 2012 to 2017, with missing values from five different locations (stations) in the State of Kuwait. 85

4.4 Time series of weather climatology (temperature and relative humidity) from 2012 to 2017, with missing values from five different locations (stations) in the State of Kuwait. 86

List of Figures

4.5	Daily concentration for SO_2 , NO_2 , weather temperature and relative humidity after estimating missing values using missForest approach from 2012-2017	90
4.6	Daily concentration for O_3 , CO , weather temperature and relative humidity after estimating missing values using missForest approach from 2012-2017	91
4.7	The steps of implementing multiple imputations using Rubin's rules to estimate missing values for the Kuwait Registry for Rheumatic Diseases (KRRD)	98
6.1	The location of monitoring fixed stations selected for this study (ASA, FAH, JAH and MAN) that belong to Kuwait Environment Public Authority (K-EPA)	161
6.2	RA patient visit form; the information in this form will be used for calculating DAS28 by using the formula 6.1 for the RA patient.	164
6.3	K-EPA air quality data process and flow chart.	165
6.4	A Pearson correlation coefficient heat map between RA disease activity scores (DAS28) and the daily average concentrations of SO_2 , NO_2 , O_3 , TEMP, RH and WS. Note that * $p < .05$, ** $p < .01$, *** $p < .001$	170
6.5	The distribution performance for Air quality index (AQI) using Cullen and Frey graphs for five governorate monitoring fixed stations in Kuwait during the period from 2013 to 2020. This figure captured the distribution performance for SO_2 , NO_2 and O_3 monitored by ASA and FAH fixed stations.	173
6.6	The distribution performance for Air quality index (AQI) using Cullen and Frey graphs for five governorate monitoring fixed stations in Kuwait during the period from 2013 to 2020. This figure captured the distribution performance for SO_2 , NO_2 and O_3 monitored by JAH and MAN fixed stations.	175

List of Figures

6.7 The distribution performance for Air quality index (AQI) using Cullen and Frey graphs for five governorate monitoring fixed stations in Kuwait during the period from 2013 to 2020. This figure captured the distribution performance for SO_2 , NO_2 and O_3 monitored by MUT fixed station. . . . 176

6.8 After implementing transformation, the distribution performance for Air quality index (AQI) using Cullen and Frey graphs for five governorate monitoring fixed stations in Kuwait during the period from 2013 to 2020. This figure captured the distribution performance for SO_2 , NO_2 and O_3 monitored by ASA and FAH fixed stations. 178

6.9 After implementing transformation, the distribution performance for Air quality index (AQI) using Cullen and Frey graphs for five governorate monitoring fixed stations in Kuwait during the period from 2013 to 2020. This figure captured the distribution performance for SO_2 , NO_2 and O_3 monitored by JAH and MAN fixed stations. 179

6.10 After implementing transformation, the distribution performance for Air quality index (AQI) using Cullen and Frey graphs for five governorate monitoring fixed stations in Kuwait during the period from 2013 to 2020. This figure captured the distribution performance for SO_2 , NO_2 and O_3 monitored by MUT fixed station. 180

6.14 Descriptive statistics over time for each numerical variable in our data frame (DAS28, SO_2 , NO_2 , O_3 , Temp, RH and WS), a plot is made, shown in the left panel, showing where data exist (blue) and missing data (red). For clarity, only running sequences of ≥ 24 hours of missing. . . . 184

6.15 Multivariate time series graphs between DAS28 (grey line) with pollutant concentration line (red line) from the period from 2012 to 2020 based on monitoring station name. 189

6.16 Long-term (2012-2020) trends of SO_2 concentrations and DAS28 calculated using the smooth trend method based on the mean measurements for four locations. 190

List of Figures

6.17	Long-term (2012-2020) trends of NO_2 concentrations and DAS28 calculated using the smooth trend method based on the mean measurements for four locations.	191
6.18	Long-term (2012-2020) trends of O_3 concentrations and DAS28 calculated using the smooth trend method based on the mean measurements for four locations.	192
6.19	Forecast performance for the study time series models for ASA station. .	219
6.20	Forecast performance for the study time series models for FAH station. .	220
6.21	Forecast performance for the study time series models for MAN station. .	221
6.22	Forecast performance for the study time series models for JAH station . .	222
6.23	The results of the Impulse Response Functions (IRFs) based on back-transforming the VECM model to its level VAR representation for the ASA station dataset.	225
6.24	The results of the Impulse Response Functions (IRFs) based on back-transforming the VECM model to its level VAR representation for the FAH station dataset.	226
6.25	The results of the Impulse Response Functions (IRFs) based on back-transforming the VECM model to its level VAR representation for the MAN station dataset.	227
6.26	The results of the Impulse Response Functions (IRFs) based on back-transforming the VECM model to its level VAR representation for the JAH station dataset.	228
6.27	Roots of the cointegration matrix.	229
7.1	Log COVID-19 Kuwait daily cases, first difference, SACF and SPACF of the residuals of the multivariate model.	245
7.2	Daily time series for Log COVID-19 (Kuwait) compared with $\text{Log}(O_3)$, $\text{Log}(SO_2)$, $\text{Log}(NO_2)$, $\text{Log}(CO)$ and $\text{Log}(PM_{10})$	248
7.3	Impulse Responses	251
7.4	Forecast Error Variance Decomposition (FEVD).	252

List of Figures

7.5 Forecasting COVID-19 cases using VECM. 255

B.1 Missing values for air quality pollutants from 2012 to 2017 per fixed station. 273

B.2 Missing values for air quality pollutants from 2012 to 2017 per year. . . . 274

B.3 Missing value patterns for air quality measurements from 2012 to 2017.
Left: Frequency of missingness in each variable. **Right:** Observed missingness patterns in the data set. The least frequent occurring patterns are located at the top of the plot, with gradually increasing frequency towards the bottom. The y-axis shows the proportion of Non-Missing(Blue) and Missing(Yellow) values. 275

B.4 Distribution analysis for PM_{10} , SO_2 , O_3 , CO , and NO_2 during 2012 to 2017, according to site location in the State of Kuwait. It is very obvious that log transformation fixes the distribution shape for all pollutants. This step is very important—that is, normalizing the skewed data, such that they approximately conform to normality—in order to use them in the imputational calculation for more accurate results (Changyong et al., 2014). 276

B.5 Mean RMSE and MAE results for the Kuwait Environmental Public Authority (KEPA) data, in order to estimate missing values for SO_2 , NO_2 , CO , and O_3 after eliminating PM_{10} due to a high level of missing values. Results are shown for MCAR (**left**), MAR (**middle**), and MNAR (**right**) data. 277

B.6 Inspecting the trace line convergence levels using an iterative Markov Chain–Monte Carlo type of algorithm with respect to the imputed means and standard deviations for PM_{10} , CO , and temperature. These trace plots show the imputed value summaries for all imputed data sets with $m = 20$ after applying 10 iterations, in order to reach to the convergence level of stability. 278

List of Figures

B.7 Inspecting the trace line convergence levels using an iterative Markov Chain–Monte Carlo type of algorithm with respect to the imputed means and standard deviations for SO_2 , NO_2 , and O_3 . These trace plots show the imputed value summaries for all imputed data sets with $m = 20$ after applying 10 iterations, in order to reach to the convergence level of stability. Each colour in the graph represents an imputed data set, where the x-axis represents the number of iterations implemented during the imputational calculation and the y-axis represents the mean (**left**-side) and standard deviation (**right**-side) of the imputed values only. 279

B.8 Density plots with multiple imputations for SO_2 , NO_2 , PM_{10} , CO , and O_3 data. The blue line represents the observed data and the red lines are the density plots of the 20 imputed data sets. As we can see, in all density plots, the red lines almost match the blue line (the observed data), which is an indication of matching between the observed and imputed values. . . 280

C.1 Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for SO_2 before and after the Box-Cox transformation, it is obvious that the Box-Cox transformation enhances the normality performance for SO_2 in FAH location. 282

C.2 Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for NO_2 before and after the Box-Cox transformation, it is obvious that the Box-Cox transformation enhances the normality performance for NO_2 in FAH location. 283

C.3 Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for O_3 before and after the Lambert S transformation, it is obvious that the Lambert S transformation enhances the normality performance for O_3 in FAH location. 284

List of Figures

C.4 Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for SO_2 before and after the Yeo-Johnson transformation, it is obvious that the Yeo-Johnson transformation enhances the normality performance for SO_2 in JAH location. 285

C.5 Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for NO_2 before and after the Yeo-Johnson transformation, it is obvious that the Yeo-Johnson transformation enhances the normality performance for NO_2 in JAH location. 286

C.6 Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for O_3 before and after the Lambert S transformation, it is obvious that the Lambert S transformation enhances the normality performance for O_3 in JAH location. 287

C.7 Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for SO_2 before and after the Log transformation, it is obvious that the Log transformation enhances the normality performance for SO_2 in MAN location. 288

C.8 Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for NO_2 before and after the Yeo-Johnson transformation, it is obvious that the Yeo-Johnson transformation enhances the normality performance for NO_2 in MAN location. 289

C.9 Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for O_3 before and after the Lambert S transformation, it is obvious that the Lambert S transformation enhances the normality performance for O_3 in MAN location. 290

C.10 Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for SO_2 before and after the Yeo-Johnson transformation, it is obvious that the Log Transform transformation enhances the normality performance for SO_2 in MUT location. 291

List of Figures

C.11 Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for NO_2 before and after the Log transformation, it is obvious that the Log transformation enhances the normality performance for NO_2 in MUT location. 292

C.12 Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for O_3 before and after the Yeo-Johnson transformation, it is obvious that the Yeo-Johnson transformation enhances the normality performance for O_3 in MUT location. 293

List of Tables

1.1	Categories of air based on air quality index (AQI) and relevant information (EPA, 2009)	12
1.2	Kuwait Air quality index (AQI) values	13
2.1	Descriptive statistics of air pollutants in years (2012–2017) for the State of Kuwait.	28
2.2	Comparison between residential and industrial area.	30
2.3	Descriptive statistics of the air climatology.	30
2.4	Correlation between the pollutants—all stations.	33
3.1	RA patient visit demographic and clinical features groups by governorate—2012 to 2017.	55
3.2	Distribution of Kuwait ambient air pollution exposure using AQI during 2012–2017.	56
3.3	Correlation analysis between rheumatoid arthritis disease factors and AQI for SO_2 , NO_2 , CO, O_3 , and PM_{10}	57
3.4	Coefficients estimated by hierarchical linear model (HLM) for DAS28 (95% confidence interval in parentheses).	59
3.5	Coefficients estimated through HLM for CDAI (95% confidence interval in parentheses).	60

List of Tables

4.1 Distribution of Kuwait ambient air pollution exposure during 2012–2017. The total daily observations for ASA are $N = 1,779$; for FAH, $N = 1,820$; for JAH, $N = 1,819$; for MAN, $N = 1,777$; and, for RUM, $N = 1,811$. 25th is the lower quartile (25th percentile), 75th is upper quartile (75th percentile). SD: standard deviation. 82

4.2 Correlation analysis between weather climatology and air-pollution components SO_2 , NO_2 , O_3 , CO , and PM_{10} 83

4.3 Comparing the differences in Missing data by site using ANOVA test. From the results we conclude that all monitoring fixed stations are different in missing values amount for each pollutant except PM_{10} 83

4.4 RMSE comparison between the indexed original values and the imputed values using missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR) missingness patterns. . . . 88

4.5 Study variables with abbreviations and with the percentages of missing values for each variable. 97

4.6 Baseline patient characteristics of KRRD (2012 to 2020). In brackets is the percentage of cases or the standard deviation of the variable according to the type of the variable. 102

4.7 The mean and standard deviation for ESR, CRP, HAQ, and DAS28 from the original data set and the imputed data sets (IM). 103

4.8 Multiple regression coefficients with 95% confidence intervals (in parentheses) for predicting DAS28 using the original data set including the missing values. 104

4.9 Multiple regression coefficients with 95% confidence intervals (in parentheses) to predict DAS28 from other predictors from PMM- and RF_m -imputed data sets. 105

4.10 Multiple regression coefficients with 95% confidence intervals (in parentheses) to predict DAS28 from other predictors from kNN- and missForest-imputed data sets. 107

List of Tables

4.11 Comparison between imputation methods after we simulated 10%, 20%, and 30% missing data in the KRRD data set. The RMSE is used to highlight and select the best missing imputation method with the lowest RMSE score. 108

6.1 Descriptive statistics for study air pollutants per location in terms of AQI calculation. 169

6.2 The correlation test between DAS28, SO_2 , NO_2 , O_3 , Temp, RH and WS using Pearson’s Correlations 170

6.3 In-sample transformation efficacy measured by probability density function (Pdf) on the original samples ($n = 16,480$) after transformation using skewness results. Values close to one indicate normally transformed data. 177

6.4 OLS regression models of air pollution’s impact on DAS28 in Kuwait during the period from 2012 to 2020. 187

6.5 ADF root test with constant and with constant and trend for EPA datasets (ASA, FAH, MAN and JAH monitoring fixed stations). 193

6.6 AD-GLS root test with constant and with constant and trend for EPA datasets (ASA, FAH, MAN and JAH monitoring fixed stations). 195

6.7 KPSS test with constant and with constant and trend for EPA datasets (ASA, FAH, MAN and JAH monitoring fixed stations). 196

6.8 Phillips-Perron test with constant and with constant and trend for EPA datasets (ASA, FAH, MAN and JAH monitoring fixed stations). 197

6.9 VAR Lag order selection - ASA Station. 198

6.10 Johansen Cointegration Test for the different monitoring stations. 199

6.11 Granger causality test - Long-run Estimation Results. 202

6.12 The results of estimation and verification of the vector autoregressive model: VAR estimates to predict $Y_t = (DAS28_t, SO_{2,t}, NO_{2,t}, O_{3,t})'$ based on the information of air pollutants among the air monitoring fixed stations. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$ 204

List of Tables

6.12 The results of estimation and verification of the vector autoregressive model: VAR estimates to predict $Y_t = (DAS28_t, SO_{2,t}, NO_{2,t}, O_{3,t})'$ based on the information of air pollutants among the air monitoring fixed stations. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$ 205

6.12 The results of estimation and verification of the vector autoregressive model: VAR estimates to predict $Y_t = (DAS28_t, SO_{2,t}, NO_{2,t}, O_{3,t})'$ based on the information of air pollutants among the air monitoring fixed stations. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$ 206

6.12 The results of estimation and verification of the vector autoregressive model: VAR estimates to predict $Y_t = (DAS28_t, SO_{2,t}, NO_{2,t}, O_{3,t})'$ based on the information of air pollutants among the air monitoring fixed stations. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$ 207

6.13 VECM model with lag order equal to 7 and Maximum likelihood estimates, observations from 2013-01-08 to 2020-12-31 with cointegration rank = 3 and using an unrestricted constant. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$ 208

6.13 VECM model with lag order equal to 7 and Maximum likelihood estimates, observations from 2013-01-08 to 2020-12-31 with cointegration rank = 3 and using an unrestricted constant. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$ 209

6.13 VECM model with lag order equal to 7 and Maximum likelihood estimates, observations from 2013-01-08 to 2020-12-31 with cointegration rank = 3 and using an unrestricted constant. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$ 210

6.14 VECM estimates – "Air pollutants" model dependent variable: $\Delta \log(DAS28)_t$ for ASA location using an unrestricted constant. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$ 212

6.15 VECM estimates – "Air pollutants" model dependent variable: $D \log(DAS28)_t$ for JAH location using an unrestricted constant. 213

List of Tables

6.16 VECM model with lag order equal 7 and Maximum likelihood estimates, observations from 2013-01-08 to 2020-12-31 with Cointegration rank = 3 using an unrestricted constant and trend. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$ 215

6.16 VECM model with lag order equal 7 and Maximum likelihood estimates, observations from 2013-01-08 to 2020-12-31 with Cointegration rank = 3 using an unrestricted constant and trend. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$ 216

6.16 VECM model with lag order equal 7 and Maximum likelihood estimates, observations from 2013-01-08 to 2020-12-31 with Cointegration rank = 3 using an unrestricted constant and trend. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$ 217

6.17 Comparison of models based on several selection criteria (R-Square, MAPE, RMSE, MAE, MPE and Theil's U2). 218

7.1 Descriptive statistics for air pollutants. 240

7.2 Descriptive statistics for weather climatologies and number of COVID-19 cases (infected and death cases) in Kuwait. 241

7.3 Pearson's Correlation test between the number of COVID-19 cases and air pollutants in Kuwait during March 10, 2020, to December 31, 2020. . . 242

7.4 Regression Coefficients to estimate the influence from air pollutants toward the changes in COVID-19 daily cases. 242

7.5 ADF Root Tests with constant. 244

7.6 ADF Root Tests with constant and trend. 244

7.7 Unit root tests using KPSS with constant and trend. 245

7.8 Phillips-Perron (PP) Unit Root Test constant and trend. 245

7.9 Lag selection criterion VAR Test using constant model with the endogenous series $\text{Log}(\text{COVID-19 Kuwait})$, $\text{Log}(O_3)$, $\text{Log}(SO_2)$, $\text{Log}(NO_2)$, $\text{Log}(CO)$. 246

7.10 Lag selection criterion VAR Test using trend model with the endogenous series $\text{Log}(\text{COVID-19 Kuwait})$, $\text{Log}(O_3)$, $\text{Log}(SO_2)$, $\text{Log}(NO_2)$, $\text{Log}(CO)$. 247

List of Tables

7.11 Johansen Test for selecting the best cointegration rank (r) that reflects linear combinations of underlying series to form a stationary series for Log(COVID-19 Kuwait), Log(O_3), Log(SO_2), Log(NO_2), Log(CO) and Log(PM_{10}). The asterisk reflects the best cointegration rank which is at $r = 1$ 249

7.12 Cointegration regression for the series Log(COVID-19 Kuwait), Log(O_3) and Log(SO_2) with constant and trend. 249

7.13 Estimates from the Error Correction Model for the series Log(COVID-19 Kuwait), Log(O_3) and Log(SO_2) with constant and trend. 250

7.14 VECM Equation d_1_KWT_Cases with restricted trend, lag order = 2, cointegration rank order = 3. 250

7.15 VECM Equation d_1_KWT_Cases with restricted constant, lag order = 2, cointegration rank order = 3. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$ 251

A.1 Comparison between Industrial Stations using ANOVA test. 271

A.2 Comparison of the Residential Stations using ANOVA test. 272

List of Abbreviations

ACPA	Anti-Cyclic Citrullinated Peptide
ACR	American College of Rheumatology
ADF	Augmented Dickey-Fuller test
AIC	Akaike Information Criterion
AMP	Arbitrary Missing Data Pattern
ANA	Antinuclear Antibodies
AQI	Air Quality Index
ASA	Ali-Subah Al-Salem Monitoring Station
BPCA	Bayesian Principal Component Analysis
C_6H_6	Benzene
CDAI	Clinical Disease Activity Index
cDMARDs	Conventional Disease Modifying Anti-Rheumatic Drugs
CH_4	Methane
CO	Carbon Monoxide
COPD	Chronic Obstructive Pulmonary Disease
CRP	C-reactive Protein
DAS28	Disease Activity Score with 28 Examined Joints

EM	Expectation Maximisation with bootstrapping
eMISK	Environmental Monitoring Information System of Kuwait
ECT	Error Correction Term
EPA	Environmental Protection Agency
ESR	The Erythrocyte Sedimentation Rate (in mm/h)
FAH	Al-Fahaheel Monitoring Station
FEVD	Forecast Error Variance Decomposition
GH	The patient Global Health assessment (from 0 = best to 100 = worst)
HAQ	The Health Assessment Questionnaire
IRFs	Impulse Response Functions
JAH	Al-Jahra Monitoring Station
K-EPA	Environmental Public Authority of Kuwait
kNN	k-nearest neighbour
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
KRRD	Kuwait Registry for Rheumatic Diseases
MAE	Mean Absolute Error
MAN	Mansouriya Monitoring Station
MAPE	Mean Absolute Percentage Error
MAR	Missing at Random
MCAR	Missing Completely at Random
MI	Multiple Imputation
MICE	Multiple Imputations by the Chained Equations

missForest	The proposed iterative imputation method based on Random Forest
MMP	Monotone Missing Data Pattern
MNAR	Missing Not at Random
MTS	Multivariate time series
MUT	Al-Mutla Monitoring Station
$NMHC_s$	Nonmethane hydrocarbons
NO	Nitrogen Monoxide
NO_2	Nitrogen Dioxide
Nodules	Rheumatoid Nodules
NO_x	Nitrogen Oxides
O_3	Ozone
OLS	Ordinary Least Squares
PACI	The Public Authority for Civil Information
PaGH	The Patient Global Health Assessment (from 0 = best to 10 = worst)
PM_{10}	Particulate Matter (PM) with 10 Microns or Less in Diameter
$PM_{2.5}$	Particulate Matter (PM) with 2.5 Microns or Less in Diameter
PMM	Predictive Mean Matching
PP	Phillips and Perron test
PrGH	The care provider global health assessment (from 0 = best to 10 = worst)
QUR	Al-Qurain Monitoring Station
RA	Rheumatoid Arthritis

RF	Rheumatoid Factor
RF_m	Random Forest method
RUM	Al-Rumaithiya Monitoring Station
SAA	Saad Al-Abdullah Monitoring Station
SBC/BIC	Schwarz Bayesian Information Criterion
SICCA	Sicca Symptoms
SJC28	The Number of Swollen Joints (0-28)
SO_2	Sulphur Dioxide
SUB	Al-Shuaiba Monitoring Station
SUK	Al-Shuwaikh Monitoring Station
Temp.	Temperature
TJC28	The Number of Tender Joints (0-28)
VAR	Vector Autoregression
VARMA	Vector Autoregression Moving Average
VECM	Vector Error Correction Model
VOCs	Volatile Organic Compounds
WD	Wind Direction
WHO	World Health Organization
WS	Wind Speed

Chapter 1

Introduction

1.1 Background

Air pollution has become a significant issue in recent decades, with dire toxicological consequences for both human wellbeing and the ecosystem. Sources of pollution vary greatly and can be as small as a single cigarette butt or as a volcanic eruption. Collective sources exist as well, such as massive volumes of emissions from automotive machines and industrial processes (Ghorani-Azam et al., 2016).

Over four-fifths of the world's population live or work in pollution levels higher than the World Health Organization's (WHO's) approved standards (Brauer et al., 2012). About 3.6 million fatalities are linked to environmental air pollution, with an additional 4 million associated with residential sources (Lim et al., 2012). This worrisome statistic is expected to double by 2050, surpassing numerous commonly known causes of death (for instance, hypercholesterolemia) (Brook et al., 2017). Furthermore, air pollution has the potential to disrupt ecosystems, destroy monuments and buildings, and modify the earth's energy balance, resulting in severe climate change (Rao et al., 2017).

In the framework of a long-term scenario, assumptions on the control of air pollution must also be compatible with the fundamental issues of climate change alleviation and alteration. In such circumstances, pollution results are predicted to occur due to a variety of factors, including efficiency gains, human health and pollution-control strategies,

as well as other problems, such as energy access, climate alteration, and agricultural productivity.

Research has thoroughly demonstrated that air pollution has extensive effects on the onset of diseases, such as cancer, cardiovascular dysfunction, and respiratory illnesses (Robinson, 2005; Habre et al., 2014; Ayres et al., 2006; Saxena and Srivastava, 2020). Even brief exposure to air pollution can induce or exacerbate a variety of respiratory and other illnesses, including bronchitis, asthma, diabetes, RA, and chronic obstructive pulmonary disease (COPD). The health risks of exposure to air pollution have been a public health issue for almost 700 years (Powell, 2012). For much of this time, the majority of air pollution and health research case studies have scrutinised the impact of acute exposure over a few days, instead of chronic exposure in a span of months or years.

A cohort study is commonly used to determine the health hazards of prolonged exposure. One example comes from an Australian longitudinal cohort research that looked at the effects of air pollution on the health of a person who was first hospitalised with heart disease. They discovered that severe pollution exposure elevates readmissions to hospitals between 3 months to one year of release, with the effect being more pronounced in individuals with heart conditions (Afoakwah et al., 2020). Another study from China, investigated the severity of meteorological factors and air pollutants on daily cases of measles between 2005 and 2009 in Lanzhou City (Chengguan District). They discovered that air pollution and weather conditions had a delayed effect on the number of measles cases on a daily basis (Peng et al., 2020).

On the level of spatial-temporal air pollution modelling, Girguis et al. (2020) demonstrated the impact of utilising a spatiotemporal exposure prediction model that has three stages and introduced formal techniques of epidemiological health risk estimate correction using shared, multiplicative measurement error (SMME). They demonstrated that spatiotemporal models based on machine learning approaches are preferred for use because they produce superior general exposure approximations due to advances in accuracy and bias reduction. To fully appreciate the inferences of employing these updated

exposure models in the epidemiological setting, epidemiologists, exposure scientists, data scientists, and statisticians have to collaborate instead of working in isolation (Butland et al., 2019).

The discipline of air quality research comprises various subtopics: *a*) developing air pollution technology, such as painting for photocatalytic breakdown of nitrogen oxide (NO_x) gases from automobiles, *b*) formulating methods for estimating and observing air quality, *c*) identifying pollutant factors and sources, and their relationships, *d*) determining cause-effect mechanisms of air pollution, and *e*) forecasting of temporal and spatial variations in air concentrations. Difficulties exist within each branch of air quality research, which inevitably become even more difficult in developing countries due to institutional, budgetary, and technological constraints. The number of monitoring stations, in particular, is restricted, and constant observation over long periods of time is ineffective in many emerging cities. As a result, insufficient data is collected, processed, and interpreted, which prevents timely, location-specific actions necessary to adapt to and mitigate the effects of deteriorating air quality.

This thesis develops a multivariate time series model for estimating and measuring short- and long-term effects of air pollution on chronic diseases toward chronic diseases such as rheumatoid arthritis (RA), a chronic inflammatory condition affecting a person's joints. Additionally, this illness can affect several other body systems in certain individuals, including their eyes, skin, heart, lungs, and blood vessels.

Numerous limitations and barriers presented themselves in our attempt to reach the optimal level of accuracy for estimating a multivariate time series model to capture the cause-and-effect relationship between pollutants and their effect on human health. One of the greatest barriers that faced us was the lack of information within air quality investigations, that is to say that the available data had missing values. Monitor faults and errors, power blackouts, system crashes, pollution levels below detection levels, and filter modifications are all common causes of missing air pollution data (Imtiaz and Shah, 2008; Libasin et al., 2020; Alahamade et al., 2021; Alsaber et al., 2021b).

The second major barrier was the inability to easily compare patient data between

multiple patient registry databases. We needed a deeper understanding of patient data than a single database could provide, but due to privacy concerns, there was a lack of personally identifiable information within the databases that would allow us to link two different sets of data to the same patient. It was essential to link the data to the same patient so as to determine relevant characteristics, such as patient address or hospital location. This deep understanding helped us to generate the key column to link between the datasets.

Another barrier initially was a lack of programming knowledge. For example, knowledge of programming skills such as using Python or *R* Programming would help wrangling the data into many shapes. For example, the air pollution dataset usually provided data in terms of hourly observations, but the patient registries usually provided daily observations. To link the air pollution dataset with the patient dataset, we aggregated the data on air pollution using R programming to convert it from hourly observations to daily observations.

Many empirical studies have shown the direct relationship between air pollutants and most chronic diseases. Prospero et al. (1996), for example, demonstrated that sulphur dioxide (SO_2) in the atmosphere is generated from both natural and anthropogenic sources. Sulphur dioxide and its atmospheric derivatives (for instance, sulphuric acid) can affect the atmosphere on global, regional, and local levels, in addition to having negative health effects (for instance indirect and direct radiative forcing and acid deposition). Anthropogenic origins are thought to be responsible for more than 70% of worldwide sulphur dioxide emissions, with fossil-fuel combustion accounting for 50% of that.

Air pollution can be defined as an atmospheric condition in which pollutants can have harmful effects on humans, animals, and the environment (Rao et al., 2017). Pollutants include, but are not limited to, gases (nitrogen oxides, carbon monoxide, and sulphur oxides), radioactive substances, and particulate particles, identified based on having an aerodynamic diameter up to and including 2.5 micrometers in diameter ($PM_{2.5}$) or 10 micrometers in diameter (PM_{10}).

Air quality prediction assists in informing the public about the quality of air, in taking required safety precautions, and in alerting authorities and companies to take appropriate action, such as actions to reduce emission rates. Consequently, this will then aid in further reducing and avoiding air pollution exposure.

1.2 Ambient Air Pollution

Human activities pollute the air we breathe, the water we drink, and the soil we farm, all having negative effects on the environment. While the industrial revolution was beneficial for science, technology, and society, it was also harmful in that it resulted in the release of massive amounts of hazardous pollutants into the atmosphere. As a result, environmental degradation can now be seen as a complex, global, public health concern. This overarching issue can also be linked to economic, social, and legislative issues. Clearly, in these times, urbanisation and industrialisation are reaching unsettling, record-high levels over the world. According to the World Health Organization (WHO), anthropogenic air pollution "is one of the world's most serious public health threats, causing around 9 million fatalities each year" (Kumari et al., 2018).

The WHO defines ambient air pollution as potentially harmful pollutants emitted by industries, households, cars, and trucks. Of all of these pollutants, fine particulate matter (PM) has the greatest effect on human health. Most fine PM comes from fuel combustion from vehicles, power plants, industries, households, or biomass burning. The WHO estimates fine PM causes 25% of lung cancer deaths, 8% of COPD deaths, and 15% of ischemic heart disease and stroke.

Exposure to air pollution is a risk factor with significant health impacts (Manisalidis et al., 2020), including epidemiological risks involving the probability that a disease, injury, or infection will occur. The risk assessment of air pollution follows the air pollution pathway (Figure 1.1) from a) sources, through b) emissions, c) concentrations, d) exposures, e) doses, and finally to f) health impacts (Brusseau et al., 2019). Sources can be defined as the origin of the pollutant, which generally involve the type, quantity, and quality of fuel used. Emissions are air pollutants released from the source that are trans-

formed and transported through the environment. Concentrations are the amount of an air pollutant in a specific space and for a specific duration. Exposures are the concentrations of air pollutants inhaled and can be measured based on the pathways, durations, intensities, and frequencies of contact with the pollutant. Doses are how much of the exposure is deposited in the body. Health impacts accrue from doses. They can be acute (short-term) or chronic (long-term), and are non-specific, in that they have many risk factors. Monitoring and intervention can occur at any stage along this pathway. Monitoring health impacts provides the primary risk indicators, though control measures at this stage are often too late, further complicated by their non-specific nature. Likewise, monitoring doses is also too late in the air pollution pathway, further complicated by a poor understanding of many pollutants. In contrast, control measures and standards generally focus on sources, emissions, and concentrations, with recent efforts targeting exposures (Tsiouri et al., 2015; Nguyen and Marshall, 2018).

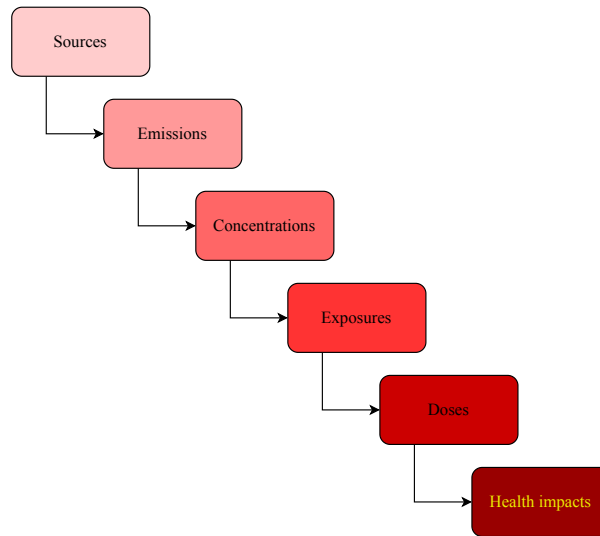


Figure 1.1: Health impact assessment following the impact pathway chain.

1.3 Fundamentals of Ambient Air Pollution

Ambient air pollution is a complex mixture of many aerosol masses— more commonly referred to as particle matter and gases. Air quality is generally measured from a small subset of these particles and gases. Two important indicators of air quality are $PM_{2.5}$ and tropospheric ozone (O_3). According to the Health Effects Institute (2018), $PM_{2.5}$ is the most consistent and most robust predictor of health effects from studies of long-term exposure to air pollution. Similarly, O_3 has been associated with increased respiratory mortality (Health Effects Institute, 2018). Quantifying exposure to ambient air pollution using $PM_{2.5}$ and O_3 as indicators is also consistent with the Global Burden of Diseases (GBD), Injuries, and Risk Factors Study (Yang et al., 2022; Karimi et al., 2019; Anenberg et al., 2018).

The 2005 WHO Air quality guidelines offer global guidance on thresholds and limits for key air pollutants that pose health risks (Organization, 2006; Organization et al., 2005). The guidelines indicate that by reducing PM_{10} pollution from 70 to 20 micrograms per cubic metre ($\mu\text{g}/\text{m}^3$), we can cut air pollution-related deaths by around 15%.

The Guidelines apply worldwide and are based on expert evaluation of current scientific evidence for:

- sulphur dioxide (SO_2),
- particulate matter (PM_{10} or $PM_{2.5}$),
- ozone (O_3), and
- nitrogen dioxide (NO_2).

1.3.1 Sulphur Dioxide (SO_2)

Historically, the main components of air pollution in many parts of the world have comprised SO_2 and PM derived from the combustion of fossil fuels (Machol and Rizk, 2013). Large urban areas have experienced the most serious problems, where coal has been used for both domestic heating purposes and poorly controlled industrial activities (Fenger,

1999; Mosley, 2014). In such situations, these sources of pollutants have generally been considered collectively, with researchers drawing on findings from epidemiological studies of heavily polluted areas completed decades earlier. To develop guidelines in this way, researchers review 24-hour averages of acute effects and annual averages of chronic effects.

Especial attention has been given to SO_2 , largely based on findings from controlled, human exposure studies. Such studies have allowed guidelines to be developed based on shorter averaging periods as short as single-hour averages which are relevant to exposures at peak concentrations that may arise from sources burning coal or heavy oil, whether or not accompanied by substantial concentrations of particulates. Epidemiological studies published in the last decade have provided further evidence on the health effects of SO_2 , warranting an independent section focusing on epidemiological results in locations mainly polluted by motor vehicles and various industries (Smith et al., 2009). A major air pollutant in many parts of the world, SO_2 derives from the combustion of sulphur-containing fossil fuels (Tsoeleng and Shikwambana, 2020). Oxidation of SO_2 , especially with metallic catalysts, leads to the formation of sulfurous acid and sulfuric acid. Neutralisation, by ammonia, leads to the production of bisulfates and sulfates. Although natural sources, such as volcanoes, contribute to environmental levels of SO_2 , in Europe, anthropogenic contributions are the greatest concern, as sulphur-containing fossil fuels are commonly burned for domestic heating and for power generation. However, in recent years the use of high-sulphur coal for domestic heating has declined in many western European countries, leaving power generation as the predominant source. This has led to a continued reduction in levels of SO_2 in cities such as London that were once heavily polluted. The use of tall chimneys at power stations has also led to widespread dispersion and dilution of SO_2 (Smith et al., 1978). These usage pattern changes have led to similar concentrations of SO_2 in urban and rural areas. In fact, in some areas, rural concentrations now exceed those in urban areas. Exposure to SO_2 causes eye-irritation and can affect the respiratory system and the functions of the lungs. Inflammation of the respiratory tract causes coughing and mucus secretion, aggravates

asthma and chronic bronchitis, and makes people more susceptible to respiratory infection. Hospital admission rates for cardiac disease and mortality rates both increase on days with higher SO_2 levels (Xu et al., 2021; Peters et al., 1999). The effects of SO_2 are not limited to those on humans. When SO_2 combines with water, it forms sulphuric acid, the main component of acid rain, which can cause deforestation.

1.3.2 Particulate Matter (PM_{10} or $PM_{2.5}$)

A common proxy indicator for air pollution is the amount of PM in the air, which affects more people than any other pollutant (Künzli et al., 2000). The major components of PM are sulfate, nitrates, ammonia, sodium chloride, black carbon, mineral dust, and water. These form a complex mixture of solid and liquid particles of both organic and inorganic substances, suspended in the air (Khaniabadi et al., 2018). While PM_{10} can penetrate and lodge deep inside the lungs, $PM_{2.5}$ are even more damaging, as they can penetrate the lung barrier and enter the blood system. Chronic exposure to PM increases one's risk of cardiovascular and respiratory diseases, as well as lung cancer (Beeson et al., 1998). Air quality measurements are typically reported in terms of daily or annual mean concentrations of PM_{10} particles per cubic metre of air volume (m^3). Routine air quality measurements typically describe such PM concentrations in terms of micrograms per cubic metre ($\mu g/m^3$). When sufficiently sensitive measurement tools are available, concentrations of fine particles ($PM_{2.5}$) are also reported. There is a close, quantitative relationship between exposure to high concentrations of PM_{10} and $PM_{2.5}$ and increased mortality or morbidity, both daily and over time (Powe and Willis, 2004). Conversely, all other factors being the same, when concentrations of small and fine particulates are reduced, related mortality rates also decrease. Understanding this relationship allows policymakers to project expected population health improvements upon reducing particulate air pollution. Small particulate pollution, of either PM_{10} or $PM_{2.5}$, has health impacts even at very low concentrations. In fact, there is no threshold below which damage has not been observed. Therefore, the WHO (2005) guidelines call for the lowest PM concentrations possible (Organization et al., 2021).

1.3.3 Ozone (O_3)

A colourless, odourless reactive gas, O_3 is comprised of three oxygen atoms. It is found naturally in the earth's stratosphere, where it absorbs the ultraviolet component of incoming solar radiation that would be harmful to life on earth. It is also found near the earth's surface, where pollutants emitted from society's activities react in the presence of sunlight to form O_3 . Principal pollutants involved in these reactions include NO_x , volatile organic compounds (VOC_s), and carbon monoxide (CO). All of these compounds are referred to as ozone precursors (Zhang et al., 2019).

Excessive O_3 in the air can have a marked effect on human health. Acute O_3 exposure can cause breathing problems, trigger asthma, reduce lung function, and cause lung diseases. Chronic exposure may cause lower lung function and deteriorated or abnormal lung development in children (Kinney et al., 2000; Gauderman et al., 2002; Zhang et al., 2019). Although the WHO also considers O_3 to be a cause of COPD, the U.S. Environmental Protection Agency (EPA) suggests there is insufficient evidence for a definitive claim (EPA, 2013). Several studies have correlated acute high O_3 concentration with increased school absences, increased visits to emergency rooms, and increased hospital admissions (Lin et al., 2008; Khorsandi et al., 2021; Malig et al., 2016; Tian et al., 2018; Niu et al., 2021). Both the WHO and the EPA consider more susceptible populations at higher risk of developing negative health effects. These include people with preexisting respiratory diseases (e.g., asthma, COPD), children, older adults, and people who are active outdoors, especially outdoor workers (WHO, 2005; EPA, 2013).

1.3.4 Nitrogen Dioxide (NO_2)

Nitrogen Dioxide (NO_2) is a specific hazardous gas among a group of highly reactive gases known as nitrogen oxides (NO_x). The primary sources of anthropogenic NO_2 emissions are combustion processes such as power generation, heating, and vehicle engines. All NO_x are harmful to human health and the environment, but NO_2 is of greatest concern for numerous reasons (Costa et al., 2014). When concentrations are $200 \mu\text{g}/\text{m}^3$ or higher, it causes extreme irritation within human airways. It is the primary source of

nitrate aerosols, which comprise up a large percentage of $PM_{2.5}$. When NO_2 is combined with ultraviolet radiation, O_3 can also be produced. Chronic exposure to NO_2 has been combined to an increase in bronchitis symptoms in children with asthma, according to epidemiological research. Evidently, studies have shown reduced lung function development to correlate with NO_2 levels currently measured in North American and European cities (Liang et al., 2016). It can also cause an increased risk of inhalation allergies. There is evidence that both the effects of chronic NO_2 exposure on mortality, and the degree of those effects, are comparable to those of $PM_{2.5}$ (Faustini et al., 2014).

1.3.5 Air Quality Index (AQI)

There are numerous air quality indices in use around the world, each with its own presentation and concept, making comparisons of air quality between other regions and cities difficult (Van den Elshout et al., 2008; Kanchan et al., 2015). The air quality index (AQI) is a numerical indicator used globally and intended to standardize the process of calculating the degree of air pollution using measured quantities of specific ambient air contaminants. In an effort to preserve the environment and human health, it informs the policymakers and public about the seriousness of air pollution and the negative effects it can have on human health. It is also used to evaluate pollution-reduction initiatives and track trends in ambient air quality (Plaia and Ruggieri, 2011). Stations monitoring ambient air quality provide a vast amount of data on a time-specific basis. These statistics are presented to stakeholders as either AQI values or alternative indices that vary in timeframe, purpose, and several other sub-indices founded on epidemiological research (Murena, 2004; Cheng et al., 2007; Kumar and Goyal, 2011; Chen et al., 2013; Van den Elshout et al., 2014; Al-Fadhli, 2017). These statistics focus on six main air pollutants: NO_2 , SO_2 , O_3 , PM_{10} , $PM_{2.5}$, and carbon monoxide (CO) (Mintz, 2006, 2009). The AQI values range from 0 to 500, and their degree is proportional to the concentration of contaminants in the ambient air, therefore a higher AQI number signals more severe potential health consequences. When the AQI exceeds 100, the air quality is considered unhealthy for sensitive groups (Organization et al., 2005). Table 1.1 shows the air pollu-

tion classification based on the AQI. A different and unique relationship occurs between AQI level and health effects for every colour-coded AQI.

Table 1.1: Categories of air based on air quality index (AQI) and relevant information (EPA, 2009)

Range of AQI	Air quality conditions/colour code	Effect on health
0-50	Good/green	Health issues do not exist because the air quality is good.
51-100	Moderate/yellow	There are no health issues because the air quality is moderate, and is considered safe for most individuals. A proportion of highly sensitive people may experience a minor effect on their health.
101-150	Unhealthy for sensitive groups/orange	Exacerbation of symptoms in persons who are prone to effects on the respiratory system and heart.
151-200	Unhealthy/red	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
201-300	Very unhealthy/purple	Exacerbation of symptoms in those who have had a heart attack or who have pulmonary illness and have had their exercise perseverance reduced; regular healthy people may experience some symptoms.
>300	Hazardous/maroon	Healthy people's exercise tolerance deteriorates, they have evident symptoms, and certain diseases manifest themselves ahead of time.

According to Johnson et al. (2010), the Air Quality Index (AQI) is defined as a measure of the condition of air relative to the requirements of one or more biotic species or to any human need (Johnson et al., 1997, 2010). The AQI is divided into categories, in which they are numbered, and each slot is marked with a colour code. This provides a scale from a healthy level of zero to a very hazardous level of above 300 as a health risk indicator associated with air quality.

The AQI is a standardized measure and a communication tool that provides a summary of ambient air quality and corresponding health risks associated with air pollution due to gases and *PM* (Kowalska et al., 2009). These indicators allow the stakeholders to track their regional, national, and local air quality without having to know the specifics of the underlying data. From the standpoint of public health, the major goal is to offer information to the policymakers and public that enables stakeholders to take necessary

steps to protect themselves from the harmful effects of air pollution. A secondary goal is to raise awareness of the effects of air pollution at existing levels of exposure, in order to motivate changes in individual behaviour as well as public policy (Doan and East, 1977; Stieb et al., 2005).

The WHO (2006) has advised nations creating policies to carefully analyse their own local situations, taking into account the unique characteristics of each location's target, namely AQI (Pruss-Ustun et al., 2006). For our research, we calculated the AQI based on the Al-Shayji et al. (2008) AQI, which was designed for the state of Kuwait based on USEPA's criteria (Fitz-Simons, 1999). In our study, Air Quality Index (AQI) is a measurement of air quality on a given day. It provides information on how clean the air is. We then, divided the values into ranges and assigned a descriptor and a colour code to each range (green for good, yellow for moderate, red for unhealthy, purple for very unhealthy, and maroon for hazardous). Every AQI range is coupled with uniform public health advice. The air quality index (AQI) varies by country and pollutant. Equation 1.1 converts each pollutant's concentration into AQI:

$$I_p = \frac{I_{high} - I_{low}}{C_{high} - C_{low}} (C_p - C_{low}) + I_{low}, \quad (1.1)$$

where I_p represents the index (Air Quality Index) for pollutant p (i.e. SO_2 , NO_2 , ..., etc.), C_p is the truncated concentration of pollutant p , C_{low} refers to the concentration breakpoint that is the less than or equal to C_p , i.e. $\leq C_p$, C_{high} refers to the concentration breakpoint that is greater than or equal to C_p , i.e. $\geq C_p$, I_{low} is the index breakpoint that is C_{low} (i.e. the AQI value corresponding to C_{low}), and I_{high} refers to the index breakpoint that is C_{high} (i.e. the AQI value corresponding to C_{high}) (Fitz-Simons, 1999).

In this study, the air quality was assessed using the AQI developed by Al-Shayji et al. (2008) for the State of Kuwait, based on the guidelines proposed by the United States Environmental Protection Agency (USEPA) (Fitz-Simons, 1999). The AQI is an index for reporting the day-to-day air quality. It gives details about the cleanliness of ambient air. Table 1.2 was used to convert from pollutant concentration to AQI:

Table 1.2: Kuwait Air quality index (AQI) values

Categories	AQI, Sub-Index	O ₃ (ppm), 8-h	PM ₁₀ (μg/m ³), 24-h	CO (ppm), 24-h	SO ₂ (ppm), 24-h	NO ₂ (ppm), 24-h
	$I_{low} - I_{high}$	$C_{low} - C_{high}$	$C_{low} - C_{high}$	$C_{low} - C_{high}$	$C_{low} - C_{high}$	$C_{low} - C_{high}$
Good	0–50	0.0–0.03	0.0–90	0.0–4.0	0.0–0.03	0.0–0.03
Moderate	51–100	0.031–0.06	90.1–350.0	4.1–8.0	0.031–0.06	0.04–0.05
Unhealthy (1)	101–150	0.061–0.092	350.1–431.1	8.1–11.7	0.061–0.182	0.06–0.30
Unhealthy (2)	151–200	0.093–0.124	431.4–512.5	11.8–15.4	0.183–0.304	0.31–0.55
Very Unhealthy	201–300	0.125–0.374	512.6–675.0	15.5–30.4	0.305–0.604	0.56–1.04
Hazardous	301–500	0.375–0.504	675.1–1000	30.5–50.4	0.605–1.004	1.05–2.04

Here we can show an example to perform the AQI calculation using equation 1.1 and table 1.2. In table 1.2, the third column shows the 24-hour PM_{10} range (low breakpoint (C_{low}) to high breakpoint (C_{high})). For "Good" air quality, This corresponds to an AQI ranging from 0 to 50. Therefore, if the 24-hour integrated PM_{10} concentration were $6.0 \mu\text{g}/\text{m}^3$ (C_p), C_{high} would be $90.0 \mu\text{g}/\text{m}^3$, C_{low} would be $0 \mu\text{g}/\text{m}^3$, circumstances I_{high} would be 50, and I_{low} would be 0. The PM_{10} range of $0 \mu\text{g}/\text{m}^3$ to $90 \mu\text{g}/\text{m}^3$ corresponds to the AQI range ($I_{low} - I_{high}$). Therefore, for a daily PM_{10} average concentration of $6.0 \mu\text{g}/\text{m}^3$, the AQI would be calculated in the following manner:

$$AQI = \frac{(50.0 - 0.0)}{(90.0 - 0.0)}(6.0 - 0.0) + 0.0 = 3.33$$

1.4 Studying the Relationship between Health and Air Pollution

The bulk of air pollution and health research focuses on the effects of acute exposure covering a few days instead of chronic exposure over years. In contrast, a cohort study is commonly used to determine the health hazards of prolonged exposure. Examples include the Harvard Six Cities Study, where researchers looked at the results of a cohort research in which nearly 8000 persons from six U.S. cities were tracked for 14-16 years (Dockery et al., 1993); the Millennium Cohort Study, which sampled roughly 19,000 babies born in Wales and England between 2000 and 2002 (Violato et al., 2009); and another that collected data on over 1.2 million individuals in 1982 (Pope et al., 1995). However, because of the large scale of sampling and the related expenditures, cohort

studies are rarely used.

As a result, time series studies account for the majority of research on the health effects of air pollution. These studies use collective-level morbidity or mortality data, which depict the health of a population residing within a given region instead of individual health. This kind of data is frequently available, making this type of study both feasible and affordable. Another benefit of time series analysis is that it is not likely to be influenced by individual-level risk variables, like smoking and age, as they are expected to remain consistent during the study period. One limitation, however, is that connections between health and exposure to air pollution from these studies can only be evaluated at the group level, which is a significantly weaker kind of analysis compared to an individual response-exposure association. This thesis will focus on a multivariate time series approach to study the relationship between air pollution and health, but it will also include a more general review of health studies and air pollution.

Data on pollution, weather, and health from a large urban region, such as a metropolis, form the basis for time series analyses. The health data consists of daily measured pollution factors and counts of morbidity and mortality outcomes for the people who live in the study area. Several fixed-site monitors have been located across the study zone and provide data that contribute to air pollution. At each site, these monitors estimate background pollution levels throughout the day and calculate a daily average for O_3 , PM_{10} , SO_2 , and NO_2 . Additionally, the fixed-site monitors routinely measure meteorological factors such as temperature, humidity, and wind speed.

1.5 Aim of the Study

The aim of this research was to build a multivariate time series model to predict the effect of air pollution using historical daily data of the state of Kuwait using multivariate time series methods, such as the vector auto-regressive (VAR) model and vector error corrected model (VECM). Descriptive statistics were collected for the AQI during the observed period. The ground-level air quality was measured both hourly and daily to determine pollutant concentration data on PM_{10} , O_3 , NO_2 , SO_2 , as well as the overall

AQI. Data were collected from the Environmental Public Authority of Kuwait (K-EPA) at a total of ten stations: seven residential and three industrial stations across Kuwait. The residential stations covered in this study included: Ali-Subah Al-Salem (SAS), Al-Fahaheel (FAH), Al-Jahra (JAH), Al-Mansouriya (MAN), Al-Qurain (QUR), Al-Rumaithiya (RUM) and Saad Al-Abdullah (SAA), and the industrial stations included Al-Mutla (MUT), Al-Shuaiba (SUB) and Al-Shuwaikh (SUK). The data corresponding to the studied pollutants were continuously monitored at these sites.

The overall goal of this research was to apply and assess commonly known time series approaches to climate change attribution and detection, as well as to use these methods to investigate causal relationships between climatic factors and chronic disease activity (i.e., RA) as dependent variable. Cointegration was used to assess Granger causality between RA and air pollution using multivariate autoregressive time series models to see if climate models could replicate actual trends. The work described in this thesis is divided into three sections, each of which focuses on a different aspect of the air pollution component of the time series model. The first and second topics are concerned with the model's ambient air pollution measurement. The bulk of studies focus on estimating and forecasting the health effects of a specific pollutant.

The questions highlighted below are addressed in this thesis:

- Can time series models provide accurate detection and attribution estimates?
- Does SO_2 Granger-cause chronic disease activity (e.g. RA)?
- Does NO_2 Granger-cause chronic disease activity (e.g. RA)?
- Does O_3 Granger-cause chronic disease activity (e.g. RA)?

In this thesis, we have implemented the multivariate time series analysis to measure the short- and long-term relationship to measure how the air pollutants predict the chronic disease activity over time series. We considered two chronic diseases, respectively to:

- patients with RA during the time period of 2013 to 2020, whose data was collected from the Kuwait Registry for Rheumatic Diseases (KRRD), and
- admitted patients with positive COVID-19 test results in Kuwait during the time period of March, 2020 to December, 2020; whose data was collected from the Ministry of Health in Kuwait (i.e. this thesis will demonstrate further evidence to establish a link between air pollution concentrations of O_3 , SO_2 , NO_2 , CO , and PM_{10} with daily COVID-19 admitted cases in the state of Kuwait.

1.5.1 Significance of Study

The ability to predict air pollution is beneficial at the macro level. This study creates value by contributing to the field of research in its evaluation of the relationship between air pollution and these diseases within Kuwait. It also aids in alerting the public about the level of air pollution in their cities so they can be aware and be careful. Predicting air pollution helps us understand how pollution affects human health, especially concerning chronic inflammatory autoimmune diseases. Without this knowledge the quality of the air is likely to be reduced, leading to respiratory problems such as lung cancer, asthma, infectious disease, and rheumatic disease.

1.5.2 The Study Region and Data

The hourly and daily air pollution datasets were collected from ten locations by the K-EPA from 1 January, 2012, to 31 December, 2020. Data on RA comes from officially registered patients of the KRRD from 1 January, 2013, to 30 December, 2020, per American College of Rheumatology (ACR) criteria (Aletaha et al., 2010; Al-Herz et al., 2016). Daily information regarding RA patient visits was collected from four main government hospitals in different Kuwait governorates, reflecting the ethnic diversity of the country's population. The Ethics Committees at Kuwait University's Faculty of Medicine and the Ministry of Health both approved the KRRD, from which this study arose. In addition, all patients who satisfied the ACR criteria for RA enrolled in the

registry gave their official consent (Al-Herz et al., 2016). The COVID-19 daily dataset was collected from the Ministry of Health in Kuwait between March and December 2020.

1.5.3 Research Objective

The primary objective of this study is to look at a newly constructed multivariate time series model that enhances air-quality forecasts in Kuwait. We want to evaluate the effectiveness of univariate and multivariate time series models, as well as earlier models, with our newly designed multivariate time series model, using the VECM technique. Six secondary objectives will help us in this task:

- **OBJECTIVE 1** Perform air-quality assessments in Kuwait from 2012 to 2017. The assessment should be matched with K-EPA's air-pollution standards.
- **OBJECTIVE 2** Deal with missing data from air pollution datasets and the other clinical dataset (KRRD). It is very important to fix and treat the missing information using advanced methods for dealing with deep-learning such as kNN and random forest to avoid biasing the results.
- **OBJECTIVE 3** Measure and test the association between disease activity scores and air pollutants, and identify the most significant pollutants contributing to the disease activity scores for RA patients.
- **OBJECTIVE 4** Dealing with variables distribution using normality assessment and developing the normality performance using transformation methods if it required. Cointegrating relationship test for unit root series to estimate the long-run (equilibrium) equation using the Engle-Granger procedure. In addition, apply the test for cointegration that allows for more than one cointegrating relationship using Johansen test to determine if three or more time series are cointegrated.
- **OBJECTIVE 5** Adopt Stationarity test by using unit root assessments such as KPSS, and Augmented Dickey-Fuller tests in order to choose the order of the model.

- **OBJECTIVE 6** Use the VECM approach to model a multivariate time series for air contaminants. Furthermore, to investigate the long-term and short-term causalities resulting from the selected air contaminants.

1.6 Structure of the Thesis

Chapter 2 explores and analyses the associations with the variables. To do this, we measured meteorological factors on the concentrations of pollutants in Kuwait using exploratory data analysis techniques. Meteorological factors included wind direction, relative humidity, wind speed, and temperature; and pollutants included O_3 , NO and NO_x , SO_2 , CO , and PM_{10} . This chapter also discusses the source of the air pollution datasets.

Chapter 3 analyses the DAS28 and CDAI indices to determine if exposure to ambient air pollution correlates with an elevated risk of RA. Specific pollutants of concern were PM_{10} , NO_2 , SO_2 , O_3 , and CO . Additionally, Chapter 3 explains the measurement of the disease activity score for the RA patients and lists their patient characteristics.

Chapter 4 treats the missing data for air pollution and RA patients by implementing several advanced methods, including the Automatic Structural Time Series Model for the air pollution dataset.

Chapter 5 discusses time series methods, cointegration, and Granger causality, with a close examination of the ramifications of non-stationarity of time series procedures and the requirement for cointegration modelling. We also use Granger causality and cointegration to show multivariate time series such as VAR and VECM. This chapter presents the rationale for the choice of data sources and variables and covers various statistical methodologies used throughout the thesis, such as the Augmented Dickey-Fuller unit root test, the Phillip-Peron unit root test, the Engle-Granger cointegration test, the Johansen cointegration test, and Granger causality tests. A further explanation is provided on how to estimate Granger causality tests using the Vector Auto Regression (VAR) model and Vector Error Correction Model (VECM) for cointegrated variables.

To address the research questions and objectives, Chapter 6 shows how to use the

Augmented Dickey-Fuller and Phillips-Perron unit root tests on out-of-sample time series data to estimate the VAR Forecast and VECM Forecast descriptive statistics, as well as their results. Chapter 7 covers the empirical results of a multivariate time series analysis of the COVID-19 patient sample data. It also presents the descriptive statistics for air pollution variables and the computation of the VAR model's optimal lag lengths, as well as the findings of the Augmented Dickey-Fuller and Phillip-Peron tests for the respective variables, the Engle-Granger and Johansen cointegration tests, and the Vector Auto Regression model's short-run causality connection (VAR). Similarly, the outcomes of the VECM's long-term causality association are discussed, as well as the short- and long-term findings of the Pairwise Granger causality tests.

Finally, Chapter 8 addresses the current study's findings in relation to the research objectives and questions, as well as its accomplishments. Other researchers may profit from the practical and theoretical outcomes of the study, thus some proposals for further research are also offered.

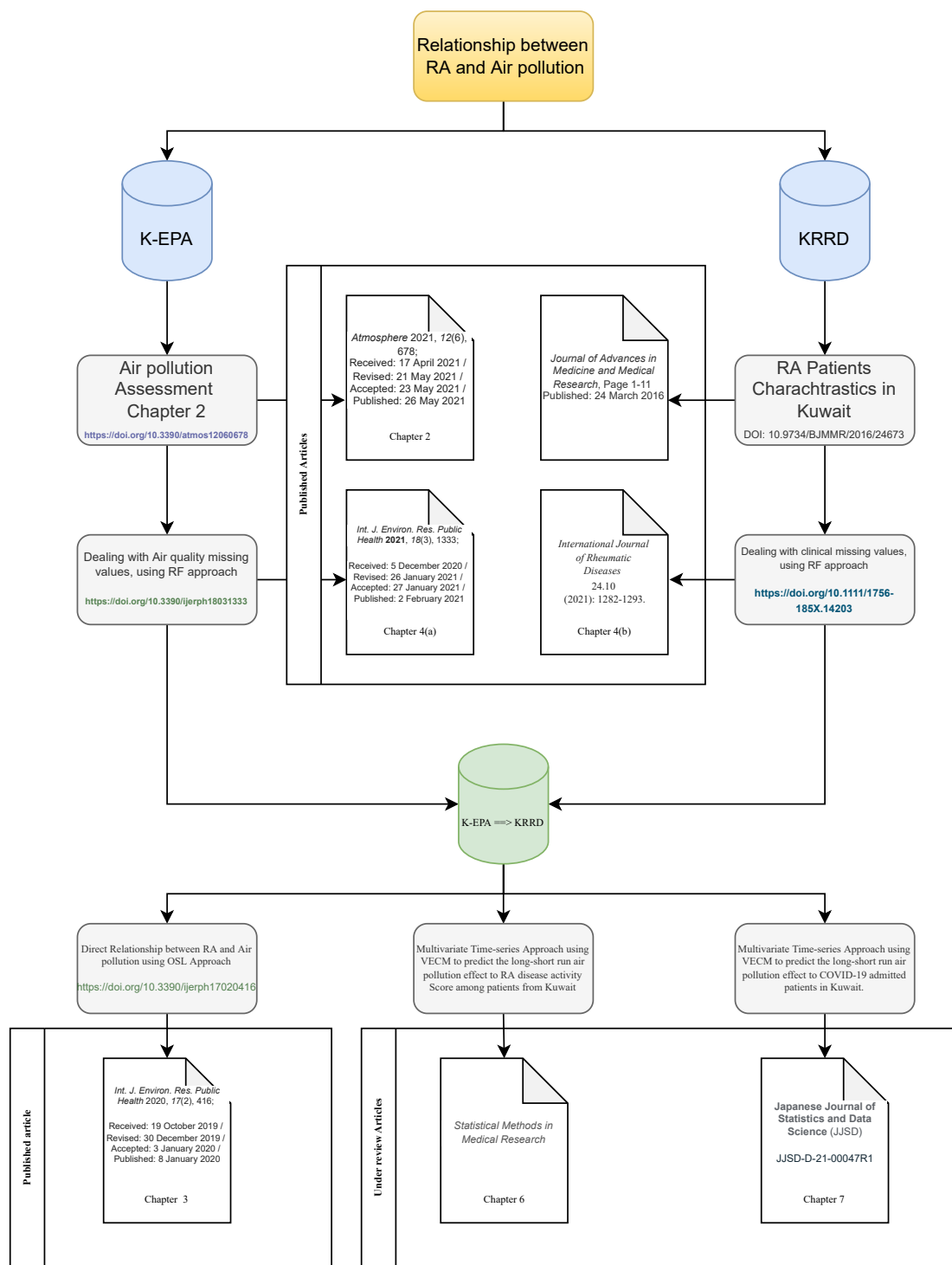


Figure 1.2: Published works from this thesis. Thesis map details can be found in this hyperlink: <https://viewer.edrawsoft.com/public/s/ae2d0447855778>

Chapter 2

Exploring the Characteristics of Air Pollution

2.1 Introduction

This chapter presents air pollution measurements from 2012 to 2017 based on ten monitoring stations at various locations across Kuwait. The monitoring stations were categorized into two distinct categories: the first category was defined as residential areas (including seven stations), and the second category was defined as industrial areas (including three stations). The main objective of this chapter is to analyse the associations with meteorological variables (wind speed (WS), wind direction (WD), temperature (Temp.) and relative humidity (RH)) on the concentrations of pollutants Ozone (O_3), Nitric oxide (NO), Nitrogen oxides (NO_x), Sulfur dioxide (SO_2), Carbon monoxide (CO), Benzene (C_6H_6), Particulate matter 10 micrometers or less in diameter (PM_{10}) and Non-methane hydrocarbons ($NMHC_s$) in Kuwait via exploratory data analysis techniques. Additionally, the pollutant concentrations mentioned previously in residential and industrial areas were compared.

2.2 Air Quality Studies

Air pollution has remained a major concern in recent decades and unfavourably affects the health of residents living in both developed and underdeveloped countries (Dockery et al., 1993; Bell and Treshow, 2002; Barnes et al., 2019). Millions of people worldwide are exposed to high levels of air pollution, which has raised human health concerns. Some of the contemporary environmental threats resulting from the consequences of human activities include greenhouse effects, ozone holes, acid rain, deforestation and photochemical smog as a main responsible threat. The combined effect of ambient (outdoor) and household (indoor) air pollution poses a major threat to health and the environment. In 2014, approximately 92% of the global population resided in areas where World Health Organization (WHO) air pollution standards were not satisfied (Birmili et al., 2014; Widiana et al., 2019). Rapid population growth and industrial development have led to an increase in pollution rates. According to the WHO, particle pollution, ground-level ozone (O_3), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), and carbon monoxide (CO) have been monitored. In addition, other pollutants occur in air comprising suspended material, such as dust, gaseous pollutants, smoke, hydrocarbons, fumes, volatile organic compounds (VOC_s), polycyclic aromatic hydrocarbons (PAH_s), and halogen derivatives, which may cause vulnerability to many diseases at high concentrations (Ghorani-Azam et al., 2016). Moreover, Alsaber et al. (2020) detected an increased risk of rheumatoid arthritis (RA) in subjects exposed to NO_2 through evaluation of the disease activity score with 28 examined joints (DAS28), and based on the Kuwait Registry for Rheumatic Diseases, they described the detrimental effects of short-term exposure to SO_2 and NO_2 on RA progression, while no correlation was found in regard to particulate matter with an aerodynamic diameter smaller than 10 microns (PM_{10}), O_3 and CO. Over the last few decades, Kuwait has experienced rapid socioeconomic and infrastructure development. The steady increase in its population, human activities, transportation fleet and power demand has contributed to environmental air pollution in Kuwait (Alenezi and Al-Anezi, 2015; Vallejo

et al., 2021). The major sources of air pollution in Kuwait include petrochemical plants, power plants, refineries and gasoline and diesel vehicles. The large number of motorized vehicles and construction expansion in industrial areas have greatly contributed to an increase in the air pollution level. In a study by Barkley et al. (2017), Kuwait was found to be the most polluted country in Southwest Asia. In July 2018, Kuwait recorded the highest air quality index (AQI) value, i.e., 301, which is hazardous and associated with serious health effects. The daily and annual concentrations of particulate matter with an aerodynamic diameter of at least 2.5 ($PM_{2.5}$) and PM_{10} in Kuwait exceeded the threshold values (daily mean $PM_{2.5}$: 10 g/m³; 24-h mean $PM_{2.5}$: 25 g/m³; daily mean PM_{10} : 20 g/m³; 24-h mean PM_{10} : 25 g/m³) defined by the WHO (Achilleos et al., 2019). Several studies on air pollution in Kuwait indicated a notable increase in various air pollutants, such as methane (CH_4), CO , O_3 , SO_2 , nitrogen oxides (NO_x) and total sulphur (TS), over a certain period (Bouhamra and Abdul-Wahab, 1999; Al-Sarawi et al., 2002; Al-Salem, 2008; Al-Mutairi et al., 2009). Another study demonstrated that traffic was the major source of air pollution in the district adjacent to the Kuwait City centre, while oil refineries contributed the most to the ambient air pollution level in a rural district (Al-Awadhi, 2014). Albassam et al. (2009) studied three pollutants, namely, CO , NO_2 and nonmethane hydrocarbons ($NMHC_s$), in the vicinity of a congested area in Kuwait. They found that the $NMHC$ concentration was much higher than the corresponding standard limit defined by the Environmental Public Authority of Kuwait (K-EPA) (an hourly maximum of 3.65 ppm and a daily average value of 1.6 ppm), which corresponded to the traffic conditions in the area. The authors focused on the impact of urban growth resulting in vehicle fleet increase in two case studies involving residential areas. They recorded excess NO_2 and $NMHC$ concentrations in both case studies. To date, no major analysis has been performed of air pollution in both industrial and residential areas, thereby identifying the sources of pollutants in Kuwait. Consequently, the aim of the present study is to measure the concentration of certain major air pollutants in industrial and residential areas. The pollutants addressed are O_3 , nitrogen monoxide (NO), NO_x , SO_2 , CO , benzene (C_6H_6), PM_{10} and $NMHC_s$, while weather variables,

such as the temperature, humidity and wind speed, were also considered.

2.3 Data and Methods

2.3.1 Description of the Study Area

The State of Kuwait is located in the northeastern corner of the Arabian Peninsula and at the top of the Arabian Gulf. It is a small developing country with a total area of 17,818 km² and depends mainly on the oil and petroleum industry. Additionally, as a desert area with a scarcity of fresh water, its main source of fresh water is desalinated sea water. Kuwait hosts three main desalination plants. Furthermore, the area is affected by severe dust storms during the summer season, which highly contribute to pollution in this area (Al-Enezi et al., 2014; Al-Ali et al., 2020). The K-EPA maintains 15 distributed air quality monitoring stations to achieve an adequate area coverage. Ten stations were selected in this study (Figure 2.1). The selection of these 10 stations was based on the observed variety of land use changes and developments, i.e., industrial and residential. This selection included the probable effect of industrial and transportation (traffic) effluents on the air quality.

2.3.2 Air Quality Data Collection

The present study is based on daily air pollutant data pertaining to the period of 2012–2017 obtained from the Environmental Public Authority at a total of ten stations: seven residential and three industrial stations across Kuwait. The residential stations covered in this study included Ali-Subah Al-Salem (ASS), Al-Fahaheel (FAH), Al-Jahra (JAH), Al-Mansouriya (MAN), Al-Qurain (QUR), Al-Rumaithiya (RUM) and Saad Al-Abdullah (SAA), and the industrial stations included Al-Mutla (MUT), Al-Shuaiba (SUB) and Al-Shuwaikh (SUK). The data corresponding to the studied pollutants were continuously monitored at these sites. The atmospheric pollutant data consisted of O_3 , NO , NO_2 , NO_x , SO_2 , CO , C_6H_6 , PM_{10} and $NMHC_s$, and the weather parameter data comprised the temperature, wind direction/speed and humidity.

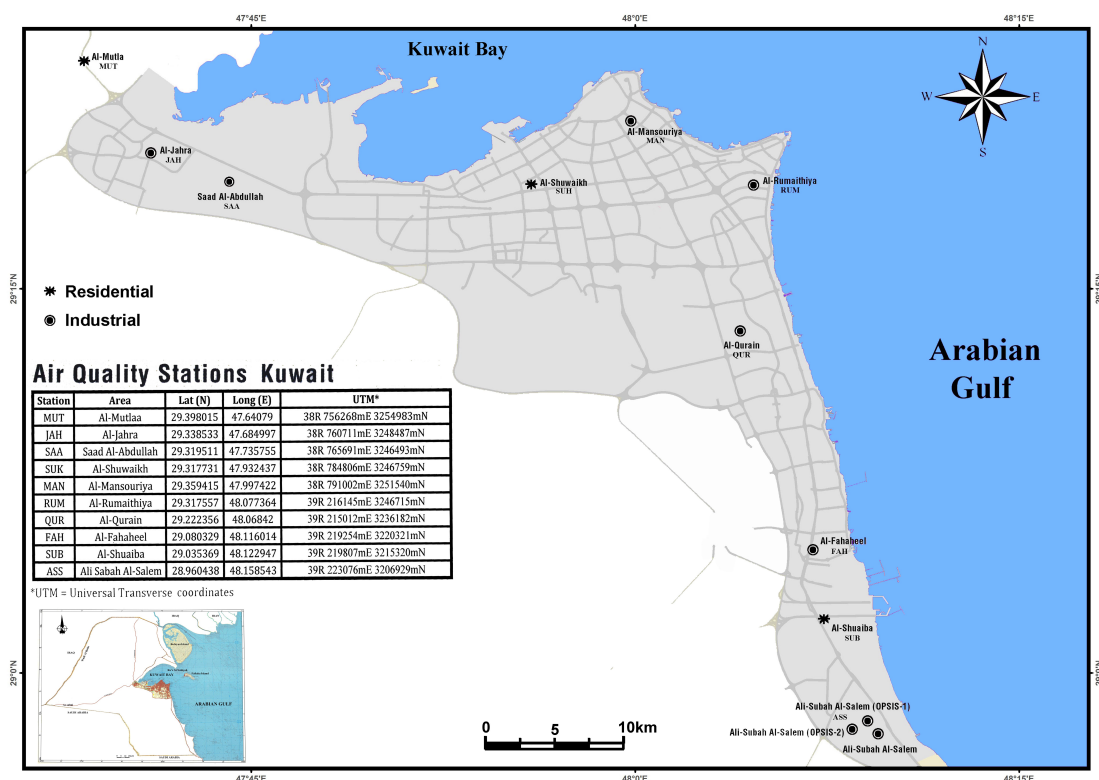


Figure 2.1: Location map of the selected monitoring stations—modified after K-EPA eMISK 2020.

K-EPA uses 15 fixed stations and 3 mobile units (Figure 2.2). According to the K-EPA method, environmental data acquisition (ENVIDAS-ENVISTA) and data transfer (every 5 min) is saved in Environmental Monitoring Information System of Kuwait (eMISK). The climatological measurements were collected at the Kuwait International Airport by the U.S. Air Force as described according to Masri et al. (2017).

2.3.3 Statistical Analysis

Descriptive analysis was employed in this study to obtain an overview of the studied variables in the form of the mean, standard deviation (S.D.), percentiles and maximum and minimum values. This represented the preliminary step to statistically analyse the different datasets. After the above descriptive analysis, correlation analysis was carried out to investigate the association among the various air pollutants and with the con-



Figure 2.2: K-EPA mobile lab and fixed stations used for air pollution monitoring.

sidered meteorological variables. In addition to correlation analysis, graphical analysis (time series, polar and box plots) was conducted to reveal the effect of meteorology and investigate the association among the addressed pollutants. Time series data are useful to extract meaningful statistics and other characteristics over time.

The data were analysed with IBM SPSS statistical software version 21 to generate descriptive statistics. Statistical data analysis was also carried out with the R programming language (R-development team, 2012) and its packages openair (Carslaw and Ropkins, 2012), ggplot2 (Wickham, 2009) and mcgv (Wood, 2003).

2.4 Air Pollution Results in Kuwait

Table 2.1 summarises the results of the descriptive statistics of the individual pollutants (O_3 , NO_2 , NO_x , NO , SO_2 , CO , C_6H_6 , PM_{10} and $NMHC_s$) over the six-year study period (2012–2017), including the average, S.D., percentiles, and maximum and minimum values. The results indicated that the average concentrations of air pollutants O_3 , NO_2 , NO_x and NO during the 2012–2017 study period were $0.02 \pm 0.01(S.D.)$, $0.03 \pm 0.02(S.D.)$, $0.05 \pm 0.04(S.D.)$ and $0.02 \pm 0.03(S.D.)$, respectively, with corresponding maximum values of 0.03, 0.42, 1.03 and 1.21, respectively. Furthermore, in the Kuwait environment, the average concentrations of air pollutants CO , PM_{10} and $NMHC_s$ were $0.82 \pm 0.73(S.D.)$, $0.22 \pm 0.85(S.D.)$ and $0.55 \pm 0.72(S.D.)$, respectively, with corresponding maximum values of 68.98, 75.22 and 59.42, respectively. The average concentrations recorded for air pollutants SO_2 and C_6H_6 were $0.01 \pm 0.01(S.D.)$ and $0.001 \pm 0.002(S.D.)$, respectively, with corresponding maximum values of 0.37 and 0.05, respectively.

Table 2.1: Descriptive statistics of air pollutants in years (2012–2017) for the State of Kuwait.

Statistic	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
O_3 (ppm)	0.024	0.013	0.0002	0.015	0.030	0.257
NO_2 (ppm)	0.033	0.022	0.0002	0.018	0.042	0.419
NO_x (ppm)	0.052	0.039	0.001	0.027	0.065	1.025
NO (ppm)	0.017	0.027	0.0003	0.006	0.020	1.207
SO_2 (ppm)	0.009	0.012	0.00000	0.004	0.009	0.366
CO (ppm)	0.815	0.725	0.005	0.489	1.072	68.980
C_6H_6 (ppm)	0.001	0.002	0.00001	0.0005	0.002	0.054
PM_{10} (g/m ³)	0.222	0.852	0.002	0.084	0.223	75.216
$NMHC$ (ppm)	0.548	0.715	0.010	0.330	0.665	59.415

Table 2.2 summarises the comparison results between the industrial and residential stations corresponding to the studied pollutants. Independent sample tests were conducted to compare the mean differences between industrial and residential stations in term of pollutant concentration. We applied independent sample t-test because in large samples (200 or more) with small standard errors, when the skewness is greater than 2 in absolute value, the variable is considered to be asymmetrical about its mean, how-

ever, robust to normality is to recognise that tests which make inferences about means, or about the expected average response at certain factor levels, are generally robust to normality. Moreover, when the kurtosis is greater than or equal to 3, then the variable's distribution is markedly different than a normal distribution in its tendency to produce outliers (Ghasemi and Zahediasl, 2012; Field, 2013; Westfall and Henning, 2013). This applies to the ANOVA test as well.

The daily mean difference among most but not all air pollutants was significant, i.e., in terms of O_3 , NO_2 , NO_x , SO_2 , CO , C_6H_6 , and $NMHC_s$, which also applied to weather parameter humidity. The analysis indicated high concentrations of NO_2 , NO_x , CO , PM_{10} and $NMHC_s$ in the residential areas, whereas the daily SO_2 and C_6H_6 concentrations were high in the industrial areas. The difference in daily concentration between air pollutants NO and PM_{10} was statistically non-significant. The recorded daily average NO_2 , NO_x , CO , PM_{10} and $NMHC$ concentrations in the residential areas were $0.04 \pm 0.02(S.D.)$, $0.05 \pm 0.04(S.D.)$, $0.88 \pm 0.80(S.D.)$, $0.23 \pm 0.99(S.D.)$ and $0.59 \pm 0.45(S.D.)$, respectively, whereas the SO_2 and C_6H_6 concentrations in the industrial areas reached $0.01 \pm 0.02(S.D.)$ and $0.002 \pm 0.002(S.D.)$, respectively.

The study results demonstrated that the overall daily average SO_2 and NO_x concentrations were lower than the corresponding K-EPA standard values in both the industrial and residential areas. Furthermore, the daily NO_2 concentration exceeded the K-EPA threshold value in the residential areas, while the daily PM_{10} concentration exceeded the K-EPA threshold value in both the industrial and residential areas.

Table 2.3 presents the descriptive statistics of the meteorological parameters (the wind speed, temperature and relative humidity). The results revealed that the average values of the wind speed, temperature and relative humidity during the 2012–2017 period were $2.65(S.D. = 1.43)$, $27.45(S.D. = 9.79)$ and $38.76(S.D. = 22.74)$, respectively.

Appendix A.1 provides the daily average concentration of the studied pollutants in the industrial areas. The comparison results were significant and indicated a significant difference among the air pollutants in the considered industrial areas. The daily concentrations of SO_2 , NO_2 and NO_x were lower than the K-EPA standard values defined for

Table 2.2: Comparison between residential and industrial area.

	I N = 4649	R N = 11,736	P-value	N
O_3 (ppm)	0.0235 (0.0153)	0.0242 (0.0120)	0.006	16,006
NO_2 (ppm)	0.0248 (0.0145)	0.0368 (0.0239)	<0.001	16,064
NO_X (ppm)	0.0454 (0.0401)	0.0535 (0.0379)	<0.001	12,058
NO (ppm)	0.0168 (0.0226)	0.0176 (0.0281)	0.063	15,136
SO_2 (ppm)	0.0094 (0.0167)	0.0082 (0.0091)	<0.001	15,953
CO (ppm)	0.6556 (0.4599)	0.8783 (0.7980)	<0.001	16,385
C_6H_6 (ppm)	0.0016 (0.0022)	0.0014 (0.0012)	0.001	4587
PM_{10} (g/m ³)	0.2130 (0.2776)	0.2261 (0.9931)	0.342	8720
$NMHC$ (ppm)	0.4264 (1.1460)	0.5928 (0.4518)	<0.001	14,349
WS	2.6662 (1.8385)	2.6444 (1.2339)	0.465	15,778
Temp.	27.4251 (10.1415)	27.4535 (9.6500)	0.872	15,747
RH	35.1748 (21.3329)	40.1788 (23.1217)	<0.001	15,751

Note: RH: Relative humidity, Temp.: Temperature in Celsius, WS: Wind speed

Table 2.3: Descriptive statistics of the air climatology.

Statistic	N	Mean	St. Dev.	Pctl(25)	Pctl(75)	Max
Wind Speed	15,778	2.651	1.432	1.692	3.300	22.771
Temperature	15,747	27.445	9.793	18.654	36.300	50.575
RH	15,751	38.757	22.739	19.833	53.583	199.000

Note: RH: Relative humidity, Temp.: Temperature in Celsius, WS: Wind speed

industrial areas except for the SUK site, where the daily NO_x concentration matched the K-EPA standard value of NO_x . The daily concentration of PM_{10} at all the sites exceeded the corresponding threshold value defined by the K-EPA. Additionally, the results demonstrated that the daily average humidity and wind speed were high at the SUB site, whereas the daily temperature was high at the SUK site.

Appendix A.2 lists the daily average concentration of the studied pollutants at the residential stations. ANOVA test was conducted to measure whether there is significant differences between the pollutant within the level of the monitoring station that located in the residential areas. The comparison results were significant and indicated a significant difference among the air pollutants in the considered residential areas. The daily concentrations of SO_2 , NO_2 and PM_{10} at all the sites exceeded the corresponding thresh-

old values defined by the K-EPA for residential areas except for the JAH site, where the daily concentration of NO_2 was lower than the standard value. Moreover, corresponding to the air pollutant NO_x , the average daily concentration was lower than the standard value in all the residential areas, while the standard value was nearly matched at only the FAH site. The results also demonstrated that the daily average humidity was high at the RUM site, whereas the daily temperature and wind speed were high at the SAA and FAH stations, respectively.

Values of the Pearson correlation coefficient are listed in Table 2.4, indicating the variation in each pollutant with respect to other air pollutants. If a given pollutant attains a strong correlation with other pollutants, it may thus be deduced that these pollutants most likely originate from the same emission source, while a low correlation coefficient value suggests different emission sources. The analysis results revealed a significantly high correlation between NO_2 and NO_x ($r_p = 0.84$), followed by that between NO and NO_x ($r_p = 0.59$), suggesting a notable dependence. Moreover, the determined high correlation coefficient value indicated a high possibility of the same emission sources for NO , NO_2 and NO_x .

The correlation among the remaining air pollutants was not strong, indicating a high possibility of different emission sources. However, the analysis results revealed a relatively high correlation between NO_2 and NO , since the presence of NO_2 in the air is a result of the NO oxidation reaction in the surrounding air ($r_p = 0.40$), followed by that between ozone (O_3) and temperature ($r_p = 0.38$). Ozone production accelerates at high temperatures in summer. Short-term exposure to Ozone has been linked to adverse health effects (Shen and Mickley, 2017).

The obtained values of the correlation coefficients were also significant for all the air pollutants except for the association between NO , CO and C_6H_6 and PM_{10} , and between C_6H_6 and SO_2 , which were statistically non-significant ($p > 0.05$). We can see from Table 2.4 that most of the pollutants resulted in negative correlation with atmospheric temperature and relative humidity; however, they showed variable response to seasonal variation of meteorological variables (e.g. temperature) and these results

agreed with Kayes et al. (2019). The analysis results indicated that the average daily concentration of pollutant SO_2 was below the K-EPA daily standard value of SO_2 for industrial areas (0.065 ppm), but it exceeded the allowable SO_2 range defined for residential areas (0.030 ppm). The analysis also indicated that the daily concentration of air pollutant NO_2 matched the K-EPA standard level of NO_2 (0.030 ppm), whereas in regard to PM_{10} , it exceeded the threshold value (0.09 g/m³). Additionally, the results demonstrated that the average daily concentration of this pollutant was below the K-EPA daily standard value (0.08 g/m³). CO and PM_{10} were characterized by the highest measurements, while the SO_2 and O_3 measurements were the lowest.

Table 2.4: Correlation between the pollutants—all stations.

	O_3	NO_2	NO_x	NO	SO_2	CO	C_6H_6	PM_{10}	$NMHC$	WS	WD	$Temp.$
O_3 (ppm)												
NO_2 (ppm)	-0.16 ***											
NO_x (ppm)	-0.23 ***	0.84 ***										
NO (ppm)	-0.18 ***	0.40 ***	0.59 ***									
SO_2 (ppm)	0.11 ***	0.24 ***	0.20 ***	0.11 ***								
CO (ppm)	-0.12 ***	0.20 ***	0.39 ***	0.20 ***	0.07 ***							
C_6H_6 (ppm)	-0.14 ***	0.28 ***	0.30 ***	0.15 ***	0.02	0.25 ***						
PM_{10} (g/m ³)	0.06 ***	-0.04 **	-0.09 ***	-0.02	-0.02 *	-0.01	-0.01					
$NMHC$ (ppm)	-0.10 ***	0.14 ***	0.13 ***	0.07 ***	0.02 *	0.12 ***	0.07 ***	-0.01				
WS	0.26 ***	-0.23 ***	-0.23 ***	-0.16 ***	0.09 ***	-0.13 ***	-0.16 ***	0.11 ***	-0.05 ***			
WD	0.11 ***	-0.17 ***	-0.16 ***	-0.09 ***	-0.16 ***	-0.20 ***	-0.08 ***	0.05 ***	-0.13 ***	0.20 ***		
$Temp.$	0.38 ***	-0.10 ***	-0.18 ***	-0.16 ***	0.00	-0.16 ***	-0.01	0.05 ***	-0.09 ***	0.17 ***	0.14 ***	
RH	-0.26 ***	0.04 ***	0.05 ***	0.06 ***	-0.03 ***	0.23 ***	0.12 ***	-0.04 **	0.08 ***	-0.16 ***	-0.28 ***	-0.61 ***

RH: Relative humidity, Temp.: Temperature in Celsius, WS: Wind speed, WD: Wind direction, * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Figure 2.3 shows the trend of the air pollutant components during the period from 2012–2017. The observed trend demonstrated that the concentrations of pollutants NO_2 , NO_x , NO , CO and $NMHC_s$ were the lowest from 2016–17, except pollutant NO , which exhibited an increasing trend before the beginning of 2017. Furthermore, it was observed that air pollutants NO_2 and NO_x exhibited a decreasing trend for the period from 2013–2016 and then an increasing trend in 2017. It was also found that the SO_2 concentration reached its highest level at a certain point during the period from 2014–2015. The analysis trend did not reveal a consistent pattern for all the pollutants. Figure 2.3 shows that the C_6H_6 , O_3 and SO_2 concentrations were lower than 0.005 ppm, 0.035 ppm and 0.015 ppm, respectively. C_6H_6 and PM_{10} did not reveal any trend during the period from 2014–2016 because of missing data values. It should be noted that due to the missing PM_{10} data and the importance of $PM_{2.5}$, it is preferable to replace PM_{10} with $PM_{2.5}$.

The daily, hourly, weekly and monthly mean variations in the pollutant concentration are shown in Figures 2.4–2.6. In regard to NO_x , NO and NO_2 , the two highest mean values were recorded in the months of January and December, and the lowest NO_x and NO_2 concentrations were recorded in June, whereas the NO concentration was the lowest during the period from June to July. The O_3 concentration exhibited the reverse pattern to that of NO_x , NO and NO_2 . The O_3 concentration peaked in July, and it gradually decreased thereafter until the end of the year, when the lowest O_3 concentration was

recorded in January and December.

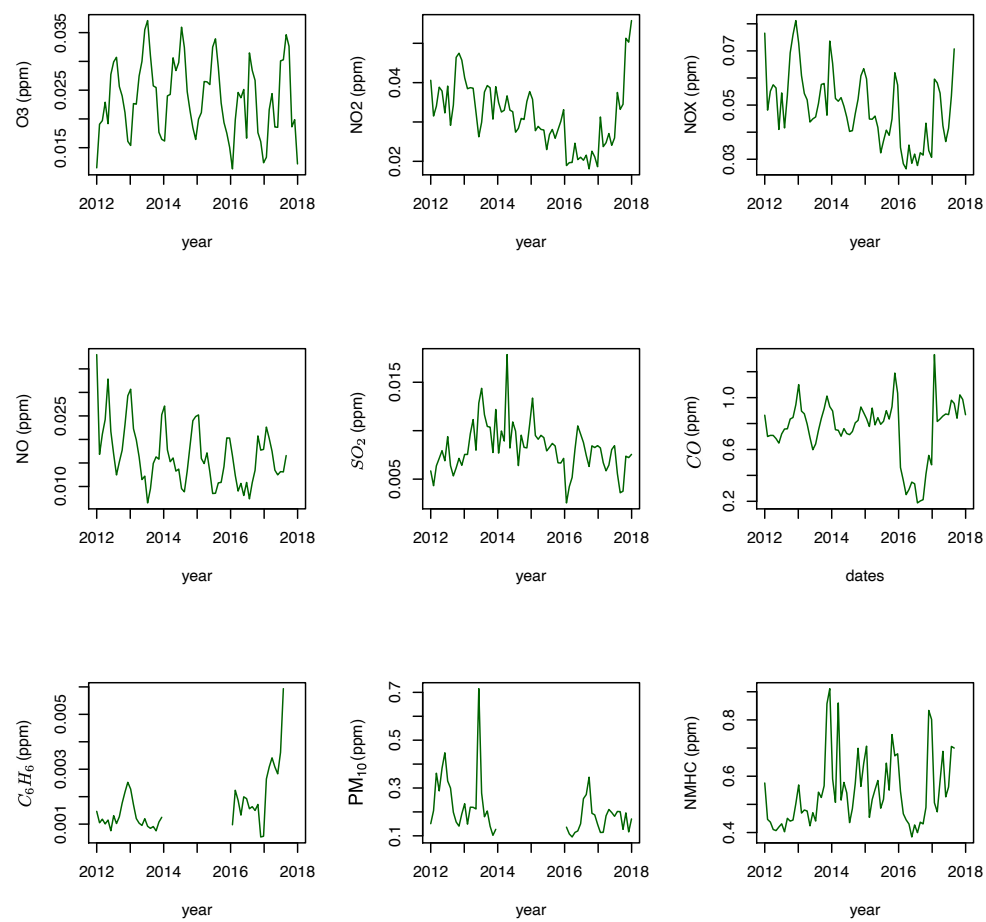


Figure 2.3: Time series of the studied pollutants from 2012 to 2017— EPA Kuwait.

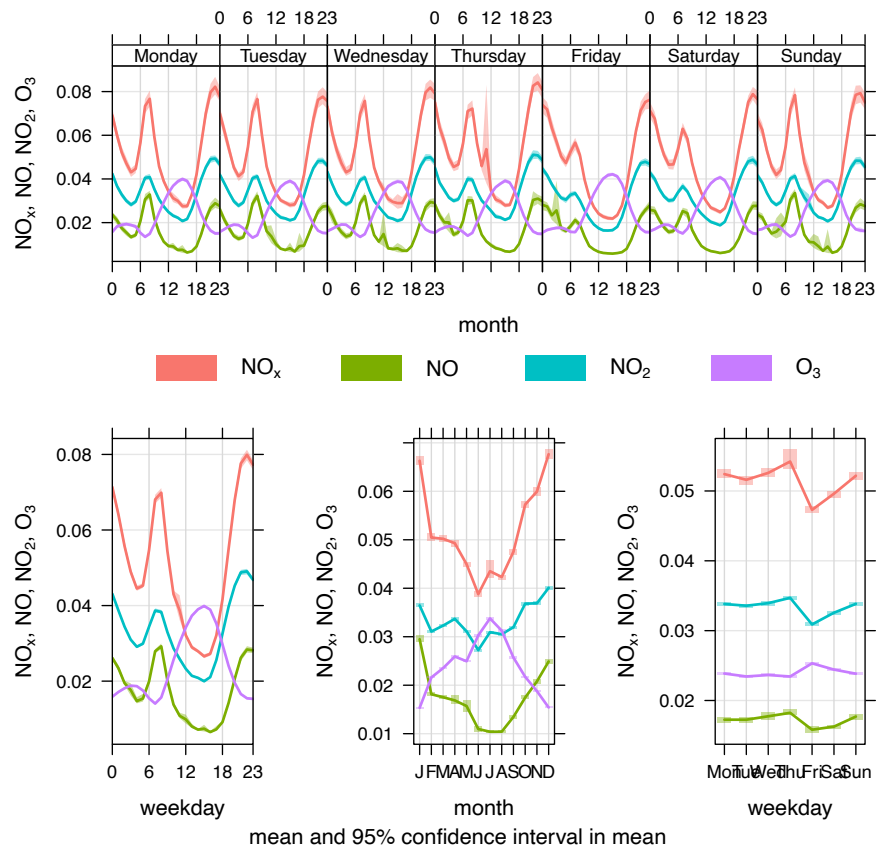


Figure 2.4: Temporal variation of the studied pollutants according to the station site from 2012 to 2017 for NO , NO_x , NO_2 and O_3 —EPA Kuwait. The shaded areas are the 95% confidence intervals for the mean. Plots created using OpenAir in R.

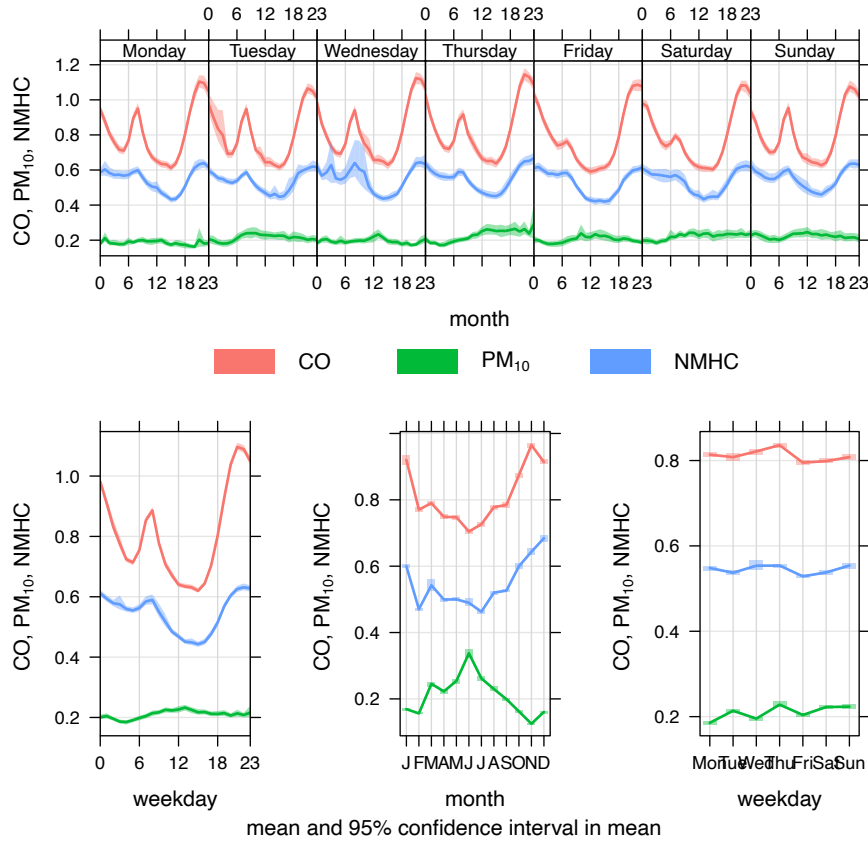


Figure 2.5: Temporal variation of the Studied Pollutants according to the station site from 2012 to 2017 for CO , PM_{10} and $NMHC$ —EPA Kuwait. The shaded areas are the 95% confidence intervals for the mean. Plots created using OpenAir in R.

Figure 2.5 shows that the concentration of pollutant CO was the highest, followed by $NMHC_s$ and PM_{10} . The figure shows that the CO and $NMHC$ concentrations were high in the winter season and low in the summer season, whereas PM_{10} exhibited the opposite trend, where the concentration was high during the summer period and low during the winter period.

Generally, regarding O_3 , a high mean concentration occurred in early summer (June and August), with low mean values observed in winter (November–February). In the present study, low nitrogen oxide emission levels (NO_x , NO and NO_2) were observed in the winter. This may occur because of the very mild temperatures in Kuwait during the winter, which led to a very low energy demand for heating purposes and resulted

in lower nitrogen oxide emission rates. However, during the summer season, a higher energy consumption was observed because of the intense and continuous use of air conditioners. A large amount of energy is required to operate this equipment, provided by the combustion of large amounts of fuel, resulting in an increase in the nitrogen oxide emission rates (NO_x , NO and NO_2).

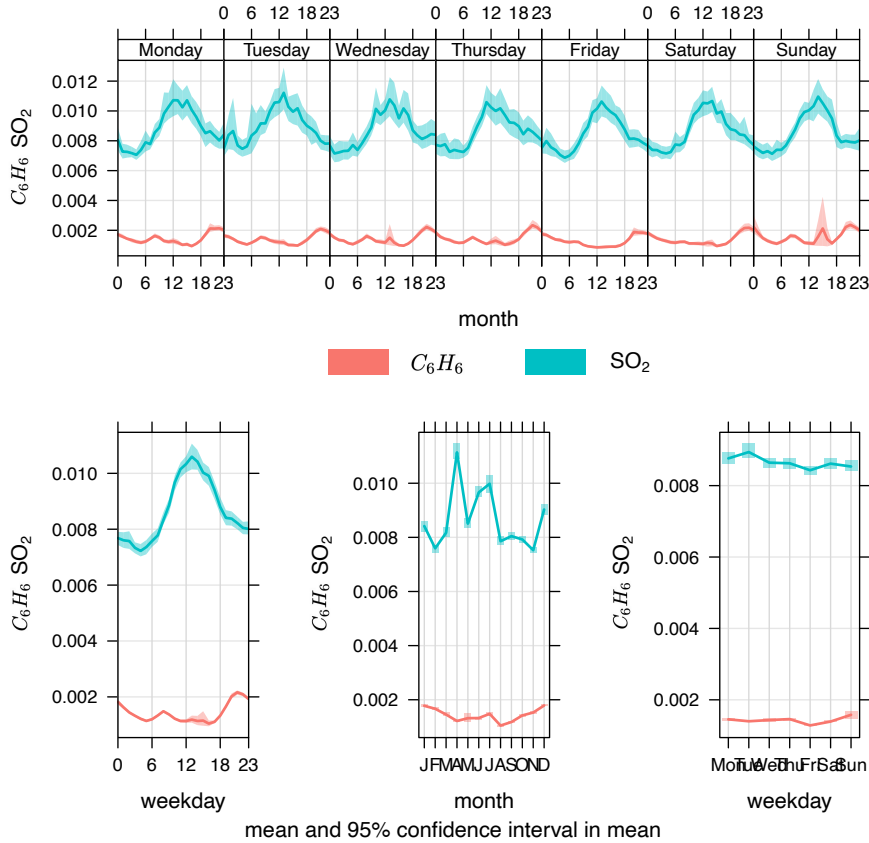


Figure 2.6: Temporal variation of the Studied Pollutants according to the station site from 2012 to 2017 for C_6H_6 and SO_2 —EPA Kuwait. The shaded areas are the 95% confidence intervals for the mean. Plots created using OpenAir in R.

Figure 2.6 shows that the SO_2 pollution level was the highest in the summer months (April and June–July), while it was the lowest in the months of February and November. The average concentration of pollutant C_6H_6 was low throughout the entire study period (2012–2017).

2.4.1 Description of Exposure Data

Box plots of the monthly pollutant concentration after suitable transformation from 2012 to 2017 are shown in figure 2.7. Box plots constitute a method to graphically depict data based on a five-number summary (minimum, first quartile (Q1), median, third quartile (Q3), and maximum). Using the daily concentration data for NO_2 , NO_x , NO , SO_2 , CO , C_6H_6 , $NMHC_s$ and PM_{10} after conducting the log transformation, however for O_3 , the square root transformation was conducted because it has been proved that ozone concentrations are most appropriately considered in terms of such a time series model on a square root transformation (Guttorp and Sampson, 1994; Carroll et al., 1997). A trend analysis was undertaken to examine the diurnal patterns and identify outliers. Figure 2.7 shows the box plots of daily emissions over 365 days and represents the median, the upper and lower quartile data range, and abnormal values shown as black circles in figure 2.7. This way it is possible to study individually the distribution of the pollutants emissions in each day. Overall, we can understand from figure 2.7 that there is variability within the hourly measurement for each pollutant and that is due to traffics with the emission of pollutants from vehicles or it might be from daily factory activity.

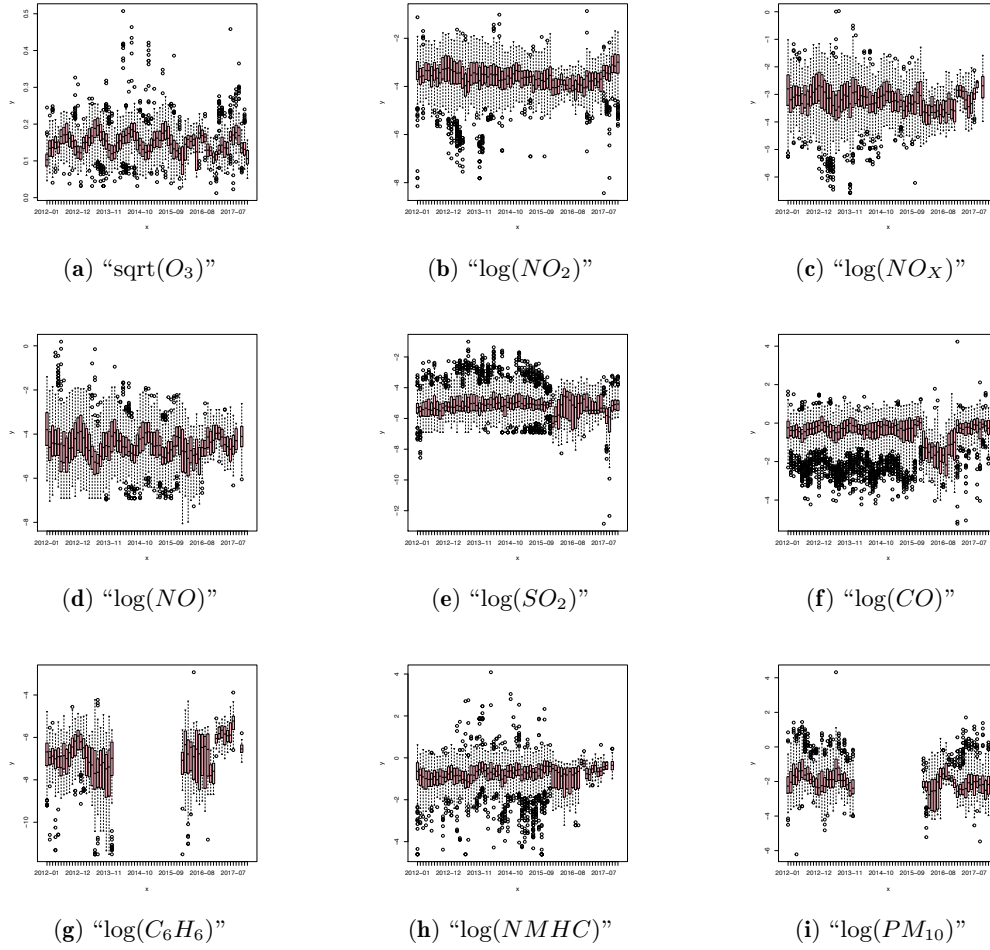


Figure 2.7: Box plot of the monthly pollutant concentration after suitable transformation from January 2012 to December 2017. The upper whisker extends to the highest value within 1.5 IQR from the top of the rectangle, while the lower whisker extends to the lowest value within 1.5 IQR from the bottom of the rectangle. Values beyond the end of the whiskers are considered outliers and are shown as dots.

Figure 2.8 shows the air pollutant concentration in the form of polar coordinates throughout the study period from 2012–2017. A polar plot shows a graphical analysis of a given database rather than a quantitative analysis. It is constructed based on the average pollutant concentration as a function of the wind speed and wind direction. Figure 2.8 shows that the concentrations of pollutants NO_2 and NO_x exhibited almost the same pattern. The concentration of these pollutants was higher at a wind speed of

5 m/s from west to east and the lowest at the northwest site. The polar plots for SO_2 , PM_{10} and $NMHC_s$ with slight variations revealed low pollutant concentrations at wind speeds ranging from 5–10 m/s. However, high SO_2 concentrations were also observed at certain points along the southeast direction. The polar plot for CO demonstrated a uniform contribution along all wind directions, except for a slightly low concentration along the east-north direction and a high concentration at a few points in time along the southeast direction at wind speeds ranging from 20–25 m/s. The high concentrations of these pollutants at low wind speeds suggested that these air pollutants may be dispersed at high wind speeds.

2.4.2 Conclusion of Air Quality Assessment in Kuwait

In the present study, time series statistical testing revealed low nitrogen oxide emission levels (NO_x , NO and NO_2) in the winter. This may occur because of the very mild temperatures in Kuwait during the winter, which led to a very low energy demand for heating purposes and resulted in lower nitrogen oxide emission rates. However, in the summer season, a higher energy consumption was observed because of the intense and continuous use of air conditioners. A large amount of energy is observed to operate air conditioners, provided by the combustion of large amounts of fuel, resulting in an increase in the nitrogen oxide emission rates (NO_x , NO and NO_2). In addition, this could be due to their locations near highways and centres of oil industries centres. Petrochemical industries and oil refineries in southern Kuwait are major sources of air pollution in the country.

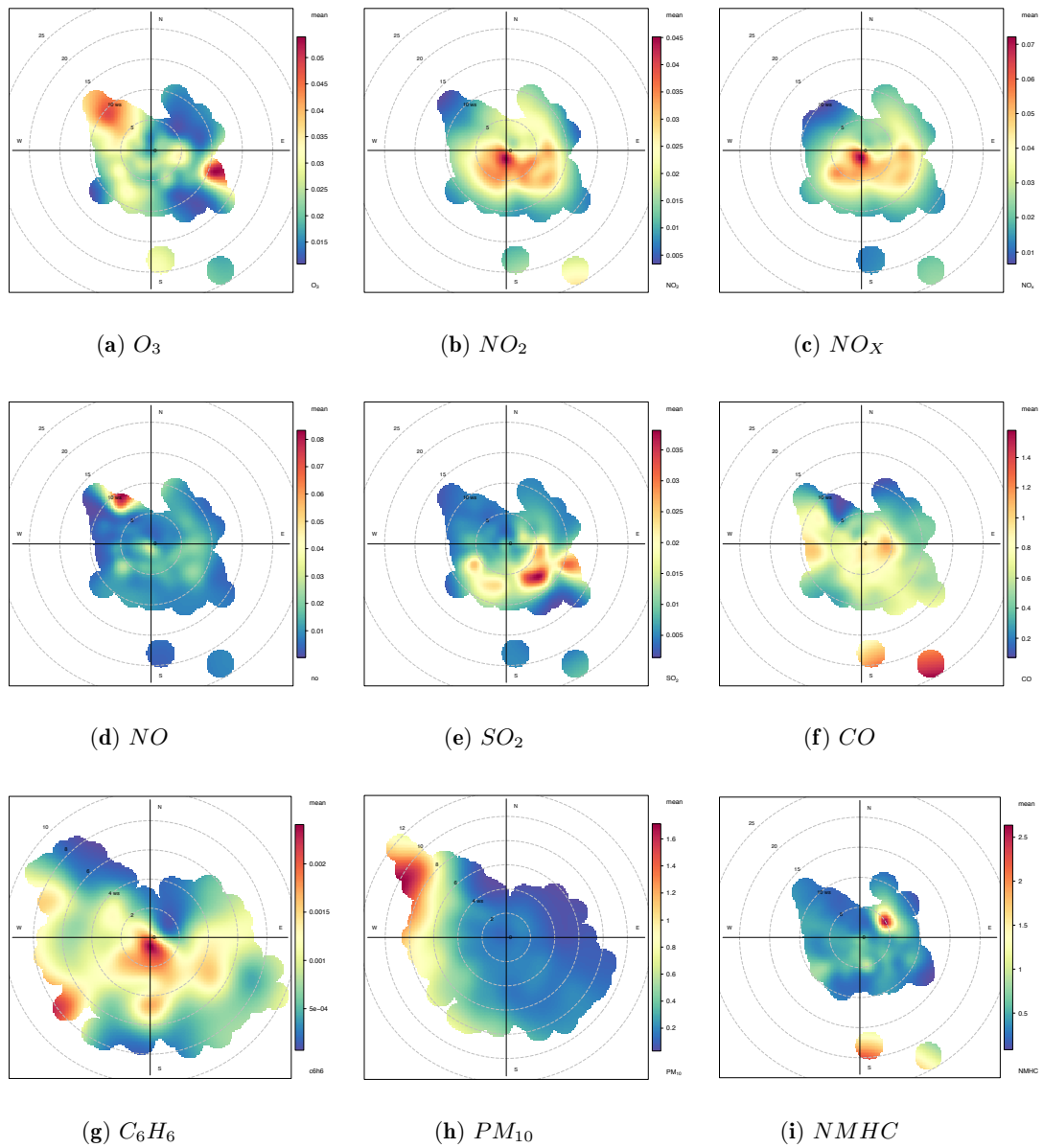


Figure 2.8: Air pollutant concentration according to the wind direction and wind speed from 2012 to 2017.

Chapter 3

The Association between the Rheumatoid Arthritis Disease Activity Score and the Ambient Air Pollution

In this chapter, we will describe the influence of ambient air pollution on the Rheumatoid Arthritis (RA) disease activity score index in 28 joints. RA is a chronic autoimmune of an unknown etiology. Air pollution has been proposed as one of the possible risk factors associated with disease activity, although has not been extensively studied. In this study, we measured the relationship between exposure to air pollutants and RA activity. Data on RA patients were extracted from the Kuwait Registry for Rheumatic Diseases (KRRD). Disease activity was measured using the disease activity score with 28 examined joints (DAS28) and the Clinical Disease Activity Index (CDAI) during a patient's hospital visits from 2013 to 2017. The assessment of DAS28 is based on the number of swollen and tender joints to provide a number/scale between 0 and 10, indicating how active the RA is at this moment (see Equation 3.2). Air pollution was assessed using air pollution components (PM_{10} , NO_2 , SO_2 , O_3 , and CO). Air pollution

data were obtained from Kuwait Environmental Public Authority (K-EPA) from six different air quality-monitoring stations during the same period. Multiple imputations by the chained equations (MICE) algorithm were applied to estimate missing air pollution data (Van Buuren et al., 2015). Patients' data were linked with air pollution data according to date and patient governorate address. Descriptive statistics, correlation analysis, and linear regression techniques were employed using STATA software. This was a cross-sectional study with a convenience sample that was performed in a rheumatology patients in Kuwait. In total, 1,651 RA patients with 9,875 follow-up visits were studied. We detected an increased risk of RA using DAS28 in participants exposed to SO_2 and NO_2 with regression coefficients $\beta = 0.003$ (95% CI: 0.0004–0.005, $p < 0.01$) and $\beta = 0.003$ (95% CI: 0.002–0.005, $p < 0.01$), respectively, but not to PM_{10} , O_3 , and CO concentrations. Conclusively, we observed a strong association between air pollution with RA disease activity. This study suggests air pollution as a risk factor for RA and recommends further measures to be taken by the authorities to control this health problem.

3.1 Introduction

Rheumatoid arthritis (RA) is a chronic autoimmune disease that mainly affects the joints, causing inflammation, pain, and difficulty to use the joints. Although the exact cause is unknown, many genetic and environmental factors have been linked to the disease. Exposure to chemicals has previously been proposed as a possible, if not the main cause of the disease (Chang et al., 2016).

Several studies suggested that exposure of air pollution may increase the risk of RA (Shen et al., 2015; Solus et al., 2015). Furthermore, epidemiological evidence indicates a significant association between the risk of RA and exposure to environmental factors, such as cigarette smoke, dioxin, noise, and traffic-related air pollution (Kobayashi et al., 2008; Hart et al., 2009; De Roos et al., 2014).

Pollutants with the strongest evidence for public health concern include *particulate*

matter (PM), ozone (O_3), nitrogen dioxide (NO_2), and sulphur dioxide (SO_2) (Organization et al., 1999). However, few studies have followed adequate methodologies correlating meteorological variables with RA. In this context, further investigation concerning the impact of air pollution on the risk of developing RA is still necessary.

With respect to RA, there is a need to investigate the impact of air pollution on RA through detailed research. Air pollutants are part of the environmental components resulting from dust storms and fossil fuel combustion, which determine RA symptoms and worsen the overall disease. Additionally, gases such as SO_2 , NO_2 , CO , and O_3 are the other main pollutants that cause RA (Sun et al., 2016). More generally, extensive investigations have been performed by several researchers about the impact of ambient air pollution on human health (Bernatsky et al., 2016; Tobón et al., 2010). For instance, SO_2 resulting from the combustion of fossil fuels with high sulphur content is considered one of the most common pollutants with the worst impact on air quality. The combustion of fossil fuels containing high sulphur content causes the release of sulphur dioxide into the atmosphere. Several researchers have adequately documented the harmful effects of long-term exposure to high levels of SO_2 on overall health (Seinfeld, 1975; Scott et al., 2003); therefore, increased emission of pollutants into the atmosphere may potentially result in several adverse health effects, including RA.

Many composite indices for RA progress measurement are actually available: for example, the Disease Activity Score (DAS) with 28 examined joints (DAS28) is one common RA index that has been extensively employed to identify the disease progress level for RA patients (Prevoe et al., 1995; Van Gestel et al., 1998; Fransen et al., 2004). The Disease Activity Score was developed to measure and assess RA disease activity in daily clinical practice, clinical trials, and long-term observational studies (Van Riel, 2014). The second RA index is the Clinical Disease Activity index (CDAI) (Smolen et al., 2003; Martins et al., 2014), which is used to assess disease activity. The CDAI was developed to provide physicians and patients with simple and more understandable instruments.

In the present study, we aimed to investigate whether exposure to ambient air pol-

lution (i.e., PM with aerodynamic diameter $< 10 \mu m$ (PM_{10}), NO_2 , SO_2 , O_3 , and CO) is associated with an increased risk of RA using the DAS28 and CDAI indices.

3.2 Materials and Methods

3.2.1 Data on RA from the Kuwait Registry for Rheumatic Diseases (KRRD)

The State of Kuwait is a small country with a total area equal to 17,818 km² located on the far west side of the Asian continent. The total population of Kuwait is around 4.6 million, distributed into six main governorates (Al-Awadhi, 2014). All RA patients represented in this study were officially registered from the Kuwait Registry for Rheumatic Diseases (KRRD) from 2013 until the end of 2017. The KRRD is a national registry listing adult patients with rheumatic diseases. Patients who fulfilled the American College of Rheumatology (ACR) criteria for RA (Aletaha et al., 2010) registered from January 2013 to December 2017 were included in the study. The RA information data were collected from the rheumatology departments of four major government hospitals in Kuwait based on patient visits. The selected hospitals are mainly distributed in different governorates covering the ethnic diversity of the Kuwaiti population. The KRRD, from which this study originated, was approved by the Ethics Committees of the Faculty of Medicine at Kuwait University, and the Ministry of Health. Additionally, official consent was obtained from all represented patients enrolled in the registry (Al-Herz et al., 2016).

3.2.2 Calculating RA Indices

RA disease activity scores are measured using two different indices: DAS28 and CDAI. The DAS28 is the sum of four outcome parameters: TJC28¹, the number of tender joints (0–28); SJC28², the number of swollen joints (0–28); ESR, the erythrocyte sedimentation rate (in mm/h) (C-reactive protein (CRP) may be used as an alternative to ESR in the calculation); and GH, the patient global health assessment (from 0 = best

¹Tender 28-Joint Count (shoulders, elbows, wrists, MCPs, PIPs including thumb IP, knees)

²Swollen 28-Joint Count (shoulders, elbows, wrists, MCPs, PIPs including thumb IP, knees)

to 100 = worst) (see Equation 3.1, (Prevoo et al., 1995)).

The second index is The Clinical Disease Activity Index (CDAI). CDAI takes into account the following items: TJC28, the number of tender joints (0–28); SJC28, the number of swollen joints (0–28); PaGH, the patient global health assessment (from 0 = best to 10 = worst); and PrGH, the care provider global health assessment (from 0 = best to 10 = worst) (see Equation 3.2, (Aletaha et al., 2005)).

$$DAS28 = 0.56 \times \sqrt{TJC28} + 0.28 \times \sqrt{SJC28} + 0.70 \times \ln(ESR \text{ Or } CRP) + 0.014 \times GH \quad (3.1)$$

$$CDAI = TJC28 + SJC28 + PaGH + PrGH \quad (3.2)$$

3.2.3 Ambient Ambient air Pollutants' Data (Environmental Public Authority of Kuwait—K-EPA)

Pollutant data (PM_{10} , NO_2 , SO_2 , O_3 , and CO) were obtained from six fixed monitoring stations run by the Environmental Public Authority of Kuwait (K-EPA). The air pollutant measurement sampling was from 1 January 2013 to 31 December 2017 based on hourly observations.

The pollutant data were distributed throughout the residential areas where stations were measuring different parameters including PM_{10} , NO_2 , SO_2 , O_3 , and CO . The hourly concentration of PM_{10} , NO_2 , and SO_2 was aggregated using the twenty-four hours, and eight-hour average concentrations of O_3 and CO . For O_3 and CO , a valid 8-hour average is one with at least 75% of the hourly data available (e.g. if there are only 6 or 7 hourly averages, divide by 6 or 7). Figure 3.1 shows the calculation of the eight-hour average concentrations of O_3 :

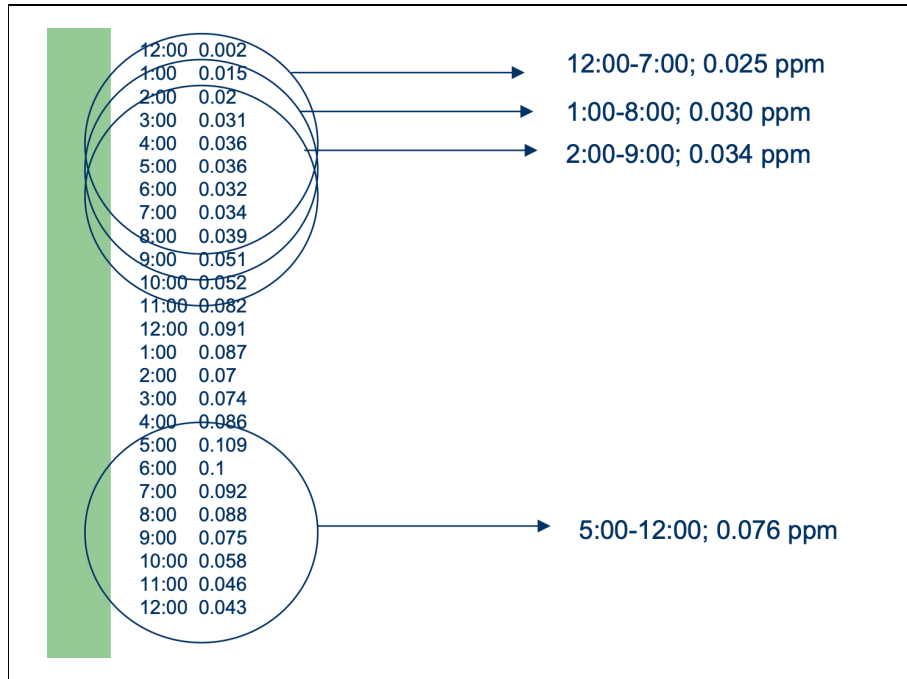


Figure 3.1: Example of eight-hour average concentrations of O_3 .

After we aggregate the air pollutants data from hourly to daily observations, all pollutants concentrations were converted from daily concentration to AQI scores. This was explained in section 1.3.5 on page 11, the AQI calculation was explained and performed using equation 1.1 on page 13 and table 1.2 on page 13.

3.2.4 Air Pollution Data Processing and Treatment

AQI data were examined by checking the normality assumption and detecting for any possible outliers before any statistical analysis or testing between the variables were done. About 5.8%, 1.6%, 48.4%, 5.3%, and 6.5% of data for PM_{10} , NO_2 , SO_2 , O_3 , and CO were missing, respectively. Multiple imputations were performed to improve the accuracy of AQI prediction, where a final estimate was composed of the outputs of several multivariate fill-in methods (Junninen et al., 2004; Schafer, 1997).

To deal with missing data, the multiple imputation process was performed to estimate and fill in missing data for better modelling performance (e.g. multiple linear regression

modelling). Finally, AQI with RA information were matched with patients' hospital location and date of patient visit. All AQI data for every pollutant were aggregated from hourly to daily observation.

3.2.5 Matching Procedure between Patients and AQI

All data management and combination was conducted using R Studio Version 1.1.463 running R 3.5.1 GUI 1.70 (Team, 2014) software. Various R packages were used to clean, match, and combine the two datasets, including *plyr* (Wickham et al., 2011), *dplyr* (Wickham and Francois, 2014), *tidyr* (Wickham, 2014), and *stringr* (Wickham, 2012).

The matching procedure was done using a developed R code to match between RA patient information and AQI monitoring station using date and governorate variables for both KRRD and K-EPA. As mentioned above, air pollution information was taken from six different stations distributed into all six governorates in the state of Kuwait. The matching procedure was conditioned on the date of the patient visit with the date of the daily average AQI using a developed R code grouped by governorate physical address (e.g., if a patient lived in the Ahmadi governorate, the AQI information that came from the Ahmadi monitoring station was added to their visit information after matching the same date; see Figure 3.2).

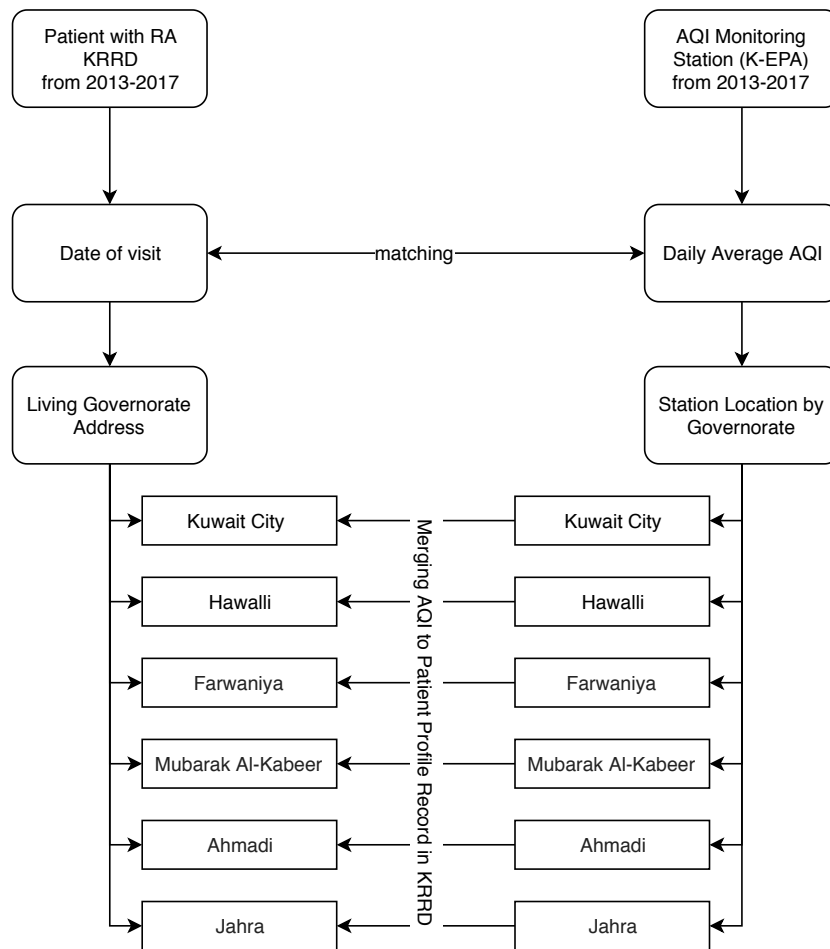


Figure 3.2: Matching procedures to combine Air Quality Index (AQI) information with Kuwait Registry for Rheumatic Diseases (KRRD) patient profile records using date and governorate address information). K-EAPL, Environmental Public Authority of Kuwait; RA, rheumatoid arthritis.

3.2.6 Statistical Analysis

In the current study, means, standard deviations (SDs), and percentages were used to summarise and compare RA characteristics between the governorate levels. To estimate the association between the pollutants and RA indices, hierarchical linear model (HLM) analysis was performed using a regression approach. HLM is a particular regression model that is designed to take into account the hierarchical or nested structure of the data. HLM is also known as multi-level modeling, linear mixed-effects model, or covari-

ance components model (Goldstein et al., 2002). With HLM we can determine the effects of potentially remediable environmental conditions (e.g., air pollution) after controlling for individual characteristics such as RA factors and demographics.

The Pearson correlation test was performed to highlight the significant associations between AQI for NO_2 , CO , PM_{10} , SO_2 , and O_3 and with the RA indices (DAS28 and CDAI). Regression analyses were performed separately for DAS28 and CDAI as response variables. Four regression models were estimated to highlight the most significant variables associated with the response variables (DAS28 or CDAI). Because the variables belong to two different databases, we started with the first regression (M1) that measured the association between patients' demographics (e.g., gender, RA disease duration, nationality, governorate, and comorbidity) with the response variables (DAS28 or CDAI), and then we estimated the second regression model (M2) that measured the association between rheumatoid factors after adjusting for gender, RA disease duration, nationality, governorate, and comorbidity with the response variables (DAS28 or CDAI). Models 1 and 2 were estimated from the KRRD database. Then, we estimated Models 3 and 4 to highlight if there is any association between air pollution with disease activity indices (DAS28 or CDAI) after merging the EPA with KRRD databases. Model 3 (M3) measured the direct effect from air pollutants to disease activity indices (DAS28 or CDAI), whereas Model 4 (M4) measured the association between air pollutants to disease activity indices after adjusting for the rheumatoid factors mentioned in Model 2 (M2) (e.g., comorbidity, treatment class, swollen, tender, etc.). For better data fit, model comparison techniques using deviance scores were implemented to confirm the best choice of model (Nelder and Wedderburn, 1972; Dobson and Barnett, 2008). All statistical procedures were performed using Stata 15.1 SE version software (StataCorp, College Station, TX, USA).

Model 1 (M1) was made to explain the influence of demographic variables (Disease Duration, Gender, Governorate, Nationality, Comorbidity, and Treatment Class) on the response variables (DAS28 and CDAI). Model 2 (M2) was made to determine the effect of RA factors (swollen, tender, RF (rheumatoid factor), anti-cyclic citrullinated

peptide (ACPA), Patient Global Assessment, and Physician Global Assessment) plus demographics on the response variables. Model 3 (M3) was made to estimate the relationship between the increase of NO_2 , CO, PM_{10} , SO_2 , and O_3 concentrations and the response variables (DAS28 and CDAI). Model 4 (M4) was made to explain the effect of AQI in terms of NO_2 , CO, PM_{10} , SO_2 , and O_3 with RA factors (swollen, RF, ACPA, ESR, and CRP) to indicate RA disease activity.

3.3 Results of RA Patient Characteristics and Air Pollution Relationship

The data of RA patients' visits were obtained from KRRD, and the air pollution data were obtained from K-EPA. The analysis was performed during the period from 2013 to the end of 2017. There were 1,651 RA patients with 9,875 follow-up visits and 13,152 daily air pollution records. Because of the matching process to combine the data from the air pollution dataset with the RA patient visits from the KRRD dataset, the final dataset had to meet the matching conditions (matching based on date and governorate) with a total of 9,875 records.

Table 3.1 shows some information about RA patient characteristics group by governorate location in the state of Kuwait. From the results, most patients were from Fawaniya governorate ($n = 4,378$ visits; 44.3%). Most of the patients belonged to the local country with Kuwaiti nationality ($n = 5,783$ visits; 58.6%). Females accounted for the majority of total visits ($n = 6,008$; 60.8%). The average RA disease duration for all patients was 9.82 years with a SD of 6.48 years.

Most of the patients were positive for rheumatoid factor (RF; $n = 6,881$; 74.6%) and positive for anti-cyclic citrullinated peptide (ACPA; $n = 4,934$; 60.5%). The majority of RA patients presented co-morbidities (e.g., hypertension, hyperlipidemia, diabetes mellitus, chronic kidney disease, coronary artery disease, cancer, or any other illness; $n = 5,393$; 54.6%). From the results, most of the consumed drugs were biologics ($n = 5,214$; 52.8%).

In Table 3.1, we can see that most RA patients visited Al-Amiri hospital, with total visits $n = 5,051$ (51.1%). Al-Amiri hospital is the only hospital in the Kuwait City governorate. The second-most patients were from Farwaniya hospital, with total visits numbering $n = 3,981$ (40.3%). Farwaniya hospital is the only public general hospital in Al Farwaniyah governorate. In Kuwait, there are six governorates, and each governorate has only one hospital.

With regard to disease activity, results in Table 3.1 present the clinical features of all RA patient visits. The average DAS28 score was 2.67 with SD 1.26, and the average score of CDAI was 6.24 with a SD of 9.96, both indicating a low disease activity. The average and SD scores for ESR and CRP were $\bar{x} = 27.19$ mm/h and $SD = 21.79$ and $\bar{x} = 6.32$ mg/L and $SD = 4.85$, respectively, which were both within the normal ranges according to our laboratory. Moreover, the average and SD for swollen and tender joints for all RA patient visits were $\bar{x} = 0.69$ and $SD = 2.26$ and $\bar{x} = 2.87$ and $SD = 5.60$, respectively.

Table 3.2 shows the average air pollutant concentrations using AQI scores. The mean and SD AQIs for PM_{10} , CO, NO_2 , O_3 , and SO_2 were 167.62 ± 214.27 , 1.28 ± 0.61 , 47.98 ± 26.64 , 17.76 ± 8.94 , and 15.87 ± 17.66 , respectively. The mean exposure levels for PM_{10} , CO, NO_2 , O_3 , and SO_2 were 158.51 ± 68.02 , 1.31 ± 0.44 , 42.74 ± 18.83 , 18.17 ± 7.56 , and 13.94 ± 12.04 , respectively. Figure 3.3 shows the monthly AQI average time series for PM_{10} , CO, NO_2 , O_3 , and SO_2 . It is very clear that the AQI for the pollutants ranged between 0 to 250. It was shown from figure 3.3 that PM_{10} ranked first in terms of pollution assessment in the State of Kuwait. It was also shown that the AQI for PM_{10} ranged between 100 to 250 which corresponds to the moderate, unhealthy and very unhealthy categories from the time series between the K-EPA monitoring fixed stations in Kuwait (see table 1.2 on page 13). And the pollutant that ranked second in terms of pollution assessment during the period from 2012 to 2017 was NO_2 . Figure 3.3 shows the AQI for NO_2 ranged between 50 to 100 during 2012 to 2014 that corresponds between moderate and unhealthy AQI categories, then after 2014, the AQI for NO_2 ranged between 0 to 50 that corresponds between good and moderate from the AQI categories. The decline in NO_2 rates was due to the government's decisions through the

K-EPA towards the industrial sector in Kuwait in the State of Kuwait to limit the NO_2 rates. For the other pollutants (CO , O_3 , and SO_2), the AQI ranged between good to moderate from the AQI categories.

Moreover, all pollutants' distributions are positively skewed (i.e., the means are higher than the medians for all pollutants). However, log transformation was employed for all pollutants in the regression model for better quality parameter estimation. Logarithmic transformation pulls extreme values in the pollutants into a more normal distribution.

In order to measure the correlation between the pollutants (PM_{10} , NO_2 , SO_2 , O_3 , and CO), the Pearson correlation test was conducted. Table 3.3 shows significant positive correlations between NO_2 and CO ($r_p = 0.22$), NO_2 and SO_2 ($r_p = 0.51$), and O_3 and PM_{10} ($r_p = 0.08$). Significant negative correlations were discovered between O_3 and NO_2 ($r_p = -0.12$), PM_{10} and NO_2 ($r_p = -0.12$), PM_{10} and SO_2 ($r_p = -0.03$), and PM_{10} and CO ($r_p = -0.05$).

Table 3.3 presents the Pearson correlation coefficients between different air pollutants and RA variables. For the score of RA disease activity using the DAS28 index, the correlation results showed a positive significant correlation with the exposure of SO_2 using AQI ($r_p = 0.07$), and the same results were returned with the exposure to NO_2 using AQI ($r_p = 0.07$). As for particular pollutants, only Hart et al. (2013b) provided evidence of elevated risks for NO_2 and SO_2 , especially in terms of seronegative RA. Other pollutants (PM_{10} , CO , and O_3) did not show any significant correlation. For the CDAI, the correlation results showed a positive significant correlation with exposure to SO_2 using AQI ($r_p = 0.10$), and the same results were returned with the exposure to NO_2 ($r_p = 0.11$). Other pollutants (PM_{10} , CO , and O_3) did not show any significant correlation with CDAI. For PM_{10} , Hart et al. (2013b) reported an elevated odds ratio (OR), which failed to reach statistical significance. The effects of both $PM_{2.5}$ and PM_{10} were non-significant across the analyses, but were consistently more pronounced for seronegative RA.

Tables 3.4 and 3.5 present the hierarchical linear model (HLM) analysis using mul-

multiple linear regression. The first model (M1) demonstrates the effect of patient demographics on RA disease activity, where DAS28 and CDAI were the response variables (see Tables 3.4 and 3.5). For both DAS28 and CDAI, the demographics weakly explained the disease activity scores ($R^2_{DAS28} = 0.034$ and $R^2_{CDAI} = 0.030$). RA disease duration was not significant for DAS28, but it had a significant association with CDAI (see Table 3.5 (M1)). Model 2 (M2) shows that RA factors (swollen, tender, RF, ACPA, patient global assessment, physician global assessment, ESR, and CRP) plus demographics affected DAS28 ($R^2 = 0.865$) and CDAI ($R^2 = 0.924$). In the CDAI model (M2), gender and RA disease duration were not significant in explaining the CDAI. Model 3 (M3) demonstrated and highlighted the effects of gaseous pollutants (PM_{10} , NO_2 , SO_2 , O_3 , and CO) using AQI on RA disease activity; only SO_2 and NO_2 were significant risk factors for RA patients using the information of DAS28 ($R^2 = 0.007$) and CDAI ($R^2 = 0.015$). The final model demonstrated the effect of gaseous air pollutants with RA factors (Swollen, RF, ACPA, ESR, and CRP) on RA disease activity. The AQI of NO_2 and SO_2 still showed positive associations with disease activity performance of RA. The positive effects of NO_2 in Model 4 (M4) were $\beta = 0.003$ (95% CI: 0.002–0.005) and $\beta = 0.048$ (95% CI: 0.030–0.066) for DAS28 and CDAI, respectively (e.g., for a $1 \mu\text{g}/\text{m}^3$ increase in daily concentration of NO_2 , DAS28 index is expected to increase by 0.003 (95%CI:0.002–0.005) and CDAI index is expected to increase by 0.048 (95% CI: 0.030–0.066), whereas, for SO_2 , the results showed a positive significant effect with $\beta = 0.003$ (95% CI: 0.0004–0.005) and $\beta = 0.044$ (95% CI: 0.018–0.070) for DAS28 and CDAI, respectively (e.g., for $1 \mu\text{g}/\text{m}^3$ increase in daily concentration of SO_2 , the DAS28 index will increase by 0.003 (95%CI: 0.0004–0.005) and the CDAI index will increase by 0.044 (95% CI: 0.018–0.070).

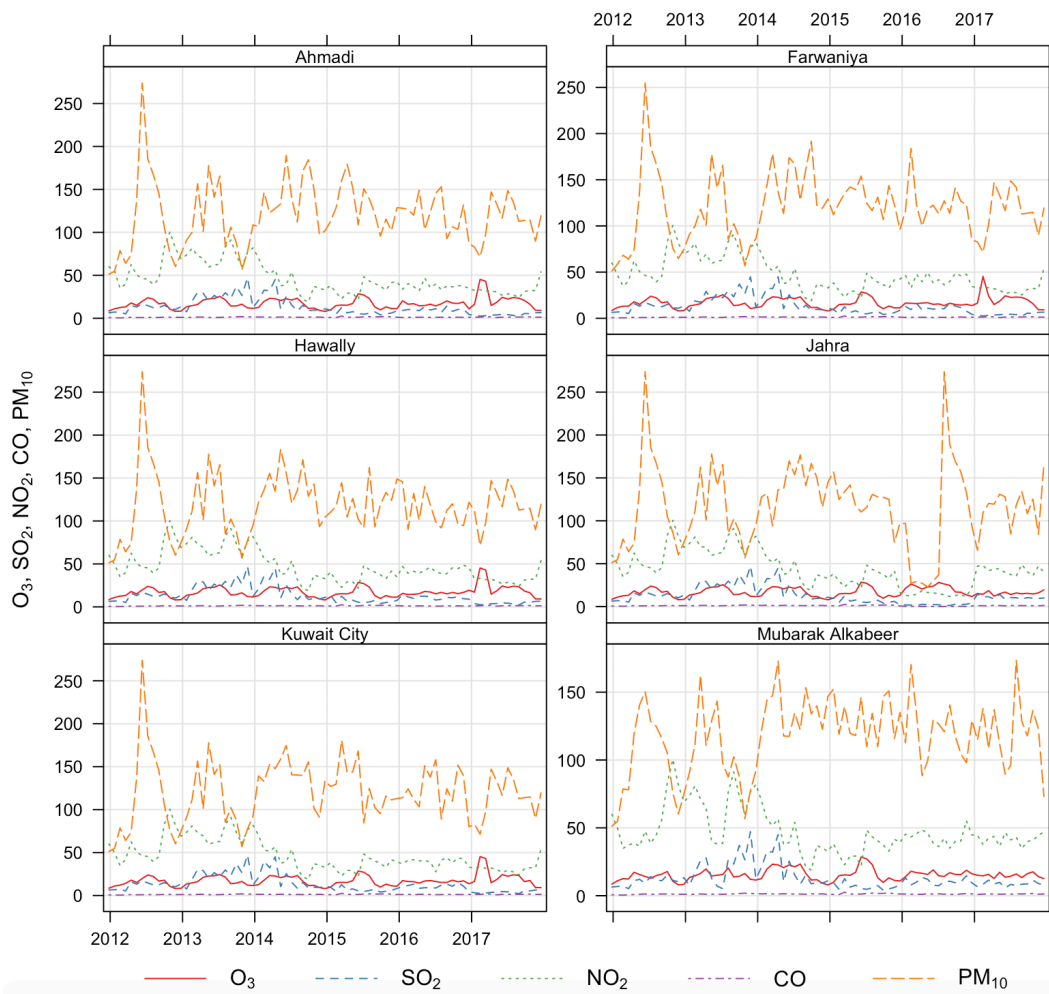


Figure 3.3: Air quality index (AQI) ambients for six governorate monitoring stations in Kuwait from 2013 to 2017.

Table 3.1: RA patient visit demographic and clinical features groups by governorate—2012 to 2017.

	[ALL]	Ahmadi	Farwaniya	Hawally	Jahra	Kuwait City	Mubarak Alkabeer
	n = 9875	n = 356	n = 4378	n = 1272	n = 226	n = 3007	n = 636
Gender:							
male	3867 (39.2%)	73 (20.5%)	2656 (60.7%)	285 (22.4%)	82 (36.3%)	653 (21.7%)	118 (18.6%)
female	6008 (60.8%)	283 (79.5%)	1722 (39.3%)	987 (77.6%)	144 (63.7%)	2354 (78.3%)	518 (81.4%)
Nationality:							
Kuwaitis	5783 (58.6%)	328 (92.1%)	1499 (34.2%)	773 (60.8%)	145 (64.2%)	2462 (81.9%)	576 (90.6%)
non-Kuwaitis	4092 (41.4%)	28 (7.87%)	2879 (65.8%)	499 (39.2%)	81 (35.8%)	545 (18.1%)	60 (9.43%)
Visited Hospital:							
Amiri	5051 (51.1%)	294 (82.6%)	364 (8.31%)	774 (60.8%)	120 (53.1%)	2955 (98.3%)	544 (85.5%)
Farwaniya	3981 (40.3%)	0 (0.00%)	3976 (90.8%)	5 (0.39%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Jahra	111 (1.12%)	0 (0.00%)	2 (0.05%)	3 (0.24%)	106 (46.9%)	0 (0.00%)	0 (0.00%)
Mubarak	732 (7.41%)	62 (17.4%)	36 (0.82%)	490 (38.5%)	0 (0.00%)	52 (1.73%)	92 (14.5%)
Disease Duration ³	9.82 (6.48)	13.3 (9.11)	9.39 (5.79)	10.4 (7.03)	10.2 (6.45)	9.42 (5.78)	11.4 (9.60)
Comorbidity ² :							
Yes	5393 (54.6%)	226 (63.5%)	1658 (37.9%)	802 (63.1%)	123 (54.4%)	2157 (71.7%)	427 (67.1%)
No	4482 (45.4%)	130 (36.5%)	2720 (62.1%)	470 (36.9%)	103 (45.6%)	850 (28.3%)	209 (32.9%)
Treatment Class:							
Biologics	5214 (52.8%)	327 (91.9%)	1342 (30.7%)	762 (59.9%)	131 (58.0%)	2124 (70.6%)	528 (83.0%)
cDMARDs ⁴	4661 (47.2%)	29 (8.15%)	3036 (69.3%)	510 (40.1%)	95 (42.0%)	883 (29.4%)	108 (17.0%)
RF ¹ :							
Positive	6881 (74.6%)	235 (72.8%)	3148 (76.3%)	817 (71.0%)	177 (93.2%)	2042 (72.0%)	462 (76.5%)
Negative	2348 (25.4%)	88 (27.2%)	976 (23.7%)	334 (29.0%)	13 (6.84%)	795 (28.0%)	142 (23.5%)
ACPA :							
Positive	4934 (60.5%)	102 (33.6%)	2665 (70.3%)	593 (63.1%)	73 (61.3%)	1205 (48.1%)	296 (60.0%)
Negative	3216 (39.5%)	202 (66.4%)	1125 (29.7%)	347 (36.9%)	46 (38.7%)	1299 (51.9%)	197 (40.0%)
Patient GA	1.64 (2.36)	1.54 (2.34)	1.02 (1.85)	2.60 (2.69)	2.09 (2.52)	1.95 (2.56)	2.39 (2.69)
Physician GA	1.05 (1.77)	1.06 (1.82)	0.72 (1.50)	1.63 (2.02)	1.58 (2.21)	1.13 (1.78)	1.64 (2.18)
DAS28	2.67 (1.26)	1.85 (1.35)	2.70 (1.21)	2.77 (1.29)	3.04 (1.39)	2.61 (1.22)	2.79 (1.44)
CDAI	6.24 (9.96)	4.64 (8.89)	4.83 (8.56)	8.31(10.72)	9.45 (14.25)	6.78 (10.49)	9.00 (11.53)
ESR	27.19 (21.79)	15.12 (16.82)	30.24 (23.10)	26.77 (20.30)	30.06 (19.33)	23.97 (19.54)	27.80 (24.04)
CRP	6.32 (4.85)	4.32 (3.91)	7.29 (4.89)	4.53 (4.53)	6.36 (4.42)	6.08 (4.68)	5.38 (4.87)
Swollen Joints	0.69 (2.26)	0.34 (1.57)	1.08 (2.60)	0.53 (1.97)	0.95 (3.63)	0.26 (1.67)	0.57 (1.99)
Tender Joints	2.87 (5.60)	1.72 (4.32)	2.02 (4.18)	3.55 (6.21)	4.82 (8.39)	3.46 (6.36)	4.54 (7.13)

¹ RF, rheumatoid factor; ACPA, anti-cyclic citrullinated peptide antibody. ² Comorbidity (e.g., hypertension, hyperlipidemia, diabetes mellitus, etc.); ³ Disease Duration, RA disease duration by years. ⁴ cDMARDs, conventional disease modifying anti-rheumatic drugs. GA, global assessment.

Table 3.2: Distribution of Kuwait ambient air pollution exposure using AQI during 2012–2017.

Air Pollutant	Ahmadi (n = 356)	Farwaniya (n = 4378)	Hawally (n = 1272)	Jahra (n = 226)	Kuwait City (n = 3007)	Mubarak Alkabeer (n = 636)	ALL (n = 9875)
PM₁₀							
min	17.108	20.346	21.948	18.837	12.138	5.421	5.421
25th ¹	75.839	76.312	80.692	67.417	71.114	74.886	74.965
median	112.623	120.779	113.586	92.788	116.081	109.917	113.586
75th ²	180.403	186.145	180.581	151.762	193.121	184.282	186.568
max	549.419	511.826	588.494	545.789	577.706	585.077	588.494
mean (SD ³)	142.36 ± 99.58	146.47 ± 94.20	145.69 ± 99.23	123.64 ± 95.00	146.81 ± 102.42	142.40 ± 102.79	144.87 ± 100.64
CO							
min	0.240	0.207	0.087	0.042	0.087	0.292	0.042
25th	0.938	0.945	0.984	0.854	1.006	0.978	0.975
median	1.367	1.338	1.337	1.151	1.369	1.380	1.346
75th	1.679	1.603	1.672	1.455	1.679	1.728	1.672
max	4.894	4.701	4.471	5.122	8.143	5.287	8.143
mean (SD)	1.37 ± 0.62	1.33 ± 0.57	1.40 ± 0.66	1.14 ± 0.64	1.40 ± 0.66	1.46 ± 0.70	1.39 ± 0.66
NO₂							
min	9.080	9.080	5.346	8.229	5.346	5.346	5.346
25th	25.768	28.200	27.208	29.127	27.319	32.037	27.391
median	35.762	36.537	35.810	52.420	36.795	44.744	37.355
75th	54.218	53.150	51.409	67.866	52.054	68.210	54.552
max	137.960	135.878	134.123	107.279	208.670	207.557	208.670
mean (SD)	42.85 ± 24.13	42.01 ± 20.70	41.29 ± 20.60	50.92 ± 25.54	43.00 ± 22.76	52.09 ± 27.91	43.74 ± 23.13
O₃							
min	4.051	4.051	4.051	5.887	3.476	4.877	3.476
25th	10.965	11.030	10.851	12.457	10.581	11.328	10.851
median	15.215	15.372	15.104	18.192	14.709	14.164	14.985
75th	20.874	22.240	20.451	25.172	20.451	19.507	20.759
max	54.262	69.682	80.656	37.539	88.623	57.330	88.623
mean (SD)	17.04 ± 8.75	18.53 ± 11.27	17.58 ± 10.52	18.71 ± 7.24	16.85 ± 10.08	16.26 ± 7.51	17.19 ± 9.90
SO₂							
min	0.003	1.000	1.000	0.665	0.003	1.000	0.003
25th	4.208	5.293	4.875	5.435	4.490	7.333	4.875
median	8.333	8.000	8.594	13.792	7.993	14.083	8.727
75th	14.146	17.292	17.169	24.583	16.746	22.946	17.504
max	121.833	121.833	121.833	76.875	111.917	127.875	127.875
mean (SD)	13.15 ± 15.46	13.26 ± 14.22	14.18 ± 16.10	17.90 ± 16.52	13.39 ± 14.60	18.62 ± 17.07	14.26 ± 15.42

¹ 25th, lower quartile (25th percentile). ² 75th, upper quartile (75th percentile). ³ SD: standard deviation.

Table 3.3: Correlation analysis between rheumatoid arthritis disease factors and AQI for SO_2 , NO_2 , CO, O_3 , and PM_{10} .

	DAS28	CDAI	NO_2	O_3	SO_2	CO	PM_{10}	Swollen	Tender	ESR
DAS28										
CDAI	0.77 ***									
NO_2	0.07 ***	0.11 ***								
O_3	0.00	0.00	-0.12 ***							
SO_2	0.07 ***	0.10 ***	0.51 ***	-0.09 ***						
CO	-0.01	0.02	0.22 ***	0.02	0.07 ***					
PM_{10}	0.00	-0.02	-0.12 ***	0.08 ***	-0.03 *	-0.05 *				
Swollen	0.50 ***	0.60 ***	0.01	0.01	0.01	0.00	-0.02			
Tender	0.72 ***	0.93 ***	0.13 ***	0.01	0.11 ***	0.03	-0.01	0.42 ***		
ESR	0.65 ***	0.20 ***	0.00	-0.02	0.04 *	-0.04 *	0.02	0.16 ***	0.17 ***	
CRP	0.28 ***	0.02 *	0.01	0.02	0.01	-0.01	-0.01	0.11 ***	0.02 *	0.37 ***

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

3.4 Discussion

Air pollution is a major concern for human health, since it is known to trigger and/or induce several pathologies and subsequently to increase morbidity and mortality rates, particularly in the Middle Eastern countries such as Kuwait. Therefore, air pollution control is a crucial element that should be prioritized by governments. According to the World Health Organization (WHO), six major air pollutants including NO_2 , CO, PM_{10} , SO_2 , O_3 , and lead (*Pb*) were identified to be primary and secondary pollutants with health risks. Many studies have found connections between particulates in the air and rates of hospitalization, chronic obstructive pulmonary disease, and restricted activity due to illness (Organization et al., 1999).

RA is considered as the most common chronic systemic auto-immune disease affecting joints, musculoskeletal apparatus, and fibrous tissues (Gabriel et al., 1999), whose incidence is expected to follow a positive trend during the following years (Chaudhari et al., 2016). In the present study, the relation between air pollutants and RA disease activity was measured using several regression models, to examine whether this association was still present even after the addition of RA factors that were highly significant for DAS28 and CDAI. From Model 1 for both DAS28 and CDAI, demographics weakly controlled the disease activity level; these results comply with patient-reported outcomes used by the National Databank for Rheumatic Diseases (NDB) (Covic et al., 2006; Godha et al., 2010).

The results of Model 1 did not show any significant evidence concerning the effect of disease duration on the disease activity level as measured by the DAS28; this also agrees with another study (Gonzalez-Alvaro et al., 2003). Nonetheless, the results of Models 3 and 4 confirmed the existence of a significant association between exposure to SO_2 and NO_2 and increase of RA disease activity: more specifically, in Model 3, where the effect of air pollutants was presented without adding RA variables, SO_2 and NO_2 showed significant relationships with RA's disease activity; in Model 4, where other RA factors were included, SO_2 and NO_2 remained risk factors for RA disease activity level.

Table 3.4: Coefficients estimated by hierarchical linear model (HLM) for DAS28 (95% confidence interval in parentheses).

	Dependent Variable			
	DAS28			
	(M1)	(M2)	(M3)	(M4)
Gender (male)	-0.213 ** (-0.268, -0.157)	-0.040 ** (-0.064, -0.017)		
RA Disease Duration	-0.002 (-0.006, 0.002)	-0.004 ** (-0.006, -0.003)		
Nationality (non-Kuwaitis)	0.272 ** (0.214, 0.331)	0.022 (-0.007, 0.051)		
Governorate (Farwaniya)	0.807 ** (0.666, 0.947)	0.299 ** (0.239, 0.358)		
Governorate (Hawally)	0.852 ** (0.704, 1.000)	0.277 ** (0.214, 0.341)		
Governorate (Jahra)	1.143 ** (0.933, 1.352)	0.254 ** (0.150, 0.359)		
Governorate (Kuwait City)	0.744 ** (0.606, 0.882)	0.290 ** (0.232, 0.348)		
Governorate (Mubarak Alkabeer)	0.955 ** (0.793, 1.117)	0.175 ** (0.106, 0.244)		
Comorbidity (Yes)	0.060 * (0.007, 0.114)	-0.051 ** (-0.074, -0.029)		
Treatment Class (cDMARDs)		0.064 ** (0.036, 0.092)		
Swollen		0.090 ** (0.085, 0.095)		0.226 ** (0.208, 0.244)
Tender		0.099 ** (0.096, 0.101)		
RF (Positive)		0.035 ** (0.010, 0.060)		0.004 (-0.078, 0.085)
ACPA (Positive)		0.008 (-0.015, 0.031)		0.007 (-0.063, 0.078)
Patient Global Assessment		0.097 ** (0.088, 0.105)		
Physician Global Assessment		0.014 * (0.002, 0.025)		
ESR		0.028 ** (0.027, 0.028)		0.035 ** (0.034, 0.037)
CRP		0.017 ** (0.015, 0.020)		0.001 (-0.006, 0.009)
NO ₂			0.003 * (0.001, 0.005)	0.003 ** (0.002, 0.005)
O ₃			0.002 (-0.002, 0.006)	0.003 (-0.001, 0.006)
SO ₂			0.004 ** (0.001, 0.007)	0.003* (0.0004, 0.005)
CO			-0.051 (-0.114, 0.012)	-0.001 (-0.053, 0.052)
PM ₁₀			0.0002 (-0.0002, 0.001)	0.00003 (-0.0003, 0.0004)
Constant	1.845 ** (1.701, 1.988)	1.029 ** (0.966, 1.093)	2.586 ** (2.435, 2.738)	1.506 ** (1.358, 1.654)
R ²	0.034	0.865	0.007	0.488
Adjusted R ²	0.033	0.865	0.006	0.486

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 3.5: Coefficients estimated through HLM for CDAI (95% confidence interval in parentheses).

	Dependent Variable			
	CDAI			
	(M1)	(M2)	(M3)	(M4)
Gender (male)	-0.748 ** (-1.187, -0.309)	-0.078 (-0.218, 0.061)		
RA Disease Duration	-0.052 ** (-0.082, -0.021)	0.005 (-0.005, 0.015)		
Nationality (non-Kuwaitis)	1.209 ** (0.747, 1.671)	0.255 ** (0.081, 0.429)		
Governorate (Farwaniya)	0.066 (-1.044, 1.176)	-1.199 ** (-1.550, -0.848)		
Governorate (Hawally)	3.408 ** (2.241, 4.576)	0.526 ** (0.152, 0.900)		
Governorate (Jahra)	4.662 ** (3.008, 6.315)	-0.559 * (-1.179, 0.062)		
Governorate (Kuwait City)	1.947 ** (0.858, 3.036)	-0.135 (-0.476, 0.207)		
Governorate (Mubarak Alkabeer)	4.430 ** (3.150, 5.709)	-0.064 (-0.474, 0.345)		
Comorbidity (Yes)	0.984 ** (0.564, 1.405)	0.147 * (0.014, 0.281)		
Treatment Class (cDMARDs)		-0.059 (-0.225, 0.107)		
Swollen		1.145 ** (1.114, 1.175)		3.035 ** (2.848, 3.222)
Tender		1.423 ** (1.410, 1.435)		
RF (Positive)		0.025 (-0.123, 0.173)		-0.066 (-0.901, 0.770)
ACPA (Positive)		0.226 ** (0.090, 0.363)		-0.594 (-1.320, 0.132)
Patient Global Assessment		1.067 ** (1.059, 1.075)		
Physician Global Assessment		0.871 ** (0.861, 0.881)		
ESR		0.019 ** (0.016, 0.022)		0.085 ** (0.068, 0.103)
CRP		-0.050 ** (-0.064, -0.037)		-0.171 ** (-0.246, -0.095)
NO ₂			0.040 ** (0.022, 0.058)	0.048 ** (0.030, 0.066)
O ₃			0.027 (-0.008, 0.062)	0.039 * (0.003, 0.074)
SO ₂			0.044 ** (0.018, 0.070)	0.044 ** (0.018, 0.070)
CO			-0.062 (-0.601, 0.476)	0.185 (-0.358, 0.729)
PM ₁₀			0.001 (-0.003, 0.004)	0.0004 (-0.003, 0.004)
Constant	4.537 ** (3.404, 5.671)	1.322 ** (0.948, 1.697)	5.306 ** (4.006, 6.606)	2.540 ** (1.017, 4.064)
	60			
R ²	0.030	0.925	0.015	0.299
Adjusted R ²	0.029	0.925	0.014	0.297

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

The Swedish Epidemiological Investigation examined the impact of prolonged exposure to air pollution on the probability of having RA: from this research, there was no evidence of a higher risk of RA resulting from exposure to PM; however, the overall risk of RA mildly increased due to atmospheric pollutants. This research confirmed the presence of a higher risk of RA following increases in NO_2 and SO_2 concentrations in ambient air (Hart et al., 2013b).

Among these various fields of research, a study using distance-to-road as a proxy component for exposure to traffic-associated pollution reported a substantial increase of RA risk in those residing within shorter distances from the road (Hart et al., 2009). Moreover, a Canadian nested case-control study reported an augmented risk of developing RA following exposure to O_3 (De Roos et al., 2014). Furthermore, other studies suggested an increased risk of RA in subjects who were exposed to NO_2 , particularly in women (Chang et al., 2016), as well as positive associations between O_3 , CO, and NO_2 and RA incidence (Jung et al., 2017).

The present study has some strengths such as combining patient records with air pollution concentration to initiate a complete dataset that could be used for future academic studies. In addition, dealing with missing values using the MICE algorithm increased the accuracy of the estimated regression coefficients. This study used the KRRD database that includes data on RA patients in Kuwait with all previous records. One of the limitations in the study is related to the records of the follow-up visits. As the data were extracted from a registry, the number of hospital visits is not equal for all patients. It ranges from 1 to 49 visits. Because of this limitation, time series analysis across patients' visits could not be performed to estimate any lag effect between air pollution and RA disease activity. Single lag day effect or moving average of several previous days' lag effect could not be investigated in this study because of the data layout and the study duration is very short (from 2013 to 2017). However, it could be developed by improving time series features in the future.

3.5 Conclusions

Air pollution was significantly associated with disease activity scores in RA patients. NO_2 and SO_2 were found to be significant risk factors for RA activity. The results confirmed that the increasing of the DAS28 can be explained by the increases of NO_2 and SO_2 with 0.7% approximate correlation as measured by the R-squared value. In addition, the increasing of the CDAI due to the increases of NO_2 and SO_2 has 1.5% approximate correlation as measured by the R-squared value for both (NO_2 and SO_2). Future research could also be based on time series analysis by employing univariate or multivariate time series analysis. It is also recommended that researchers classify the data on air pollution and disease activity score using a cluster technique and perform an adequate cluster analysis on the data.

Chapter 4

Dealing with Environmental and Clinical Missing Data

Missing data is a problem that exists within virtually any discipline that makes use of empirical data. When performing longitudinal or cross-sectional studies in any environmental research, it is not uncommon for data to be missing either by chance or by design. For instance, in research involving multiple waves of measurements, missing data can arise due to attrition, that is, subjects drop out before the end of the study.

Typically, researchers have many standard complete-data techniques available, many of which were developed early in the twentieth century like the ordinary least-squares regression and factor analysis (Seal, 1967), when there was just no solution for handling missing values. More modern techniques like the random effects model (Henderson et al., 1959) or the logistic regression (Cox, 1958) that became accessible before 1970 were also intended for complete data sets. Software packages like R, SAS, and SPSS provide these routines. However, these methods, being complete-data techniques, are not capable of dealing correctly with incomplete data sets.

Simple solutions were in use for decades (Schafer and Graham, 2002). These strategies involved discarding incomplete cases or substituting missing data by somehow plausible values. The most popular approach is complete case analysis (CCA) also known

as listwise deletion. The method is simple, and no particular modifications are needed. The main difficulty is that not all missing values have the same reason for not being observed, and there are situations in which missing data do not affect the conclusions, but generally, no justification is provided for the assumptions underlying the analysis at hand.

In this chapter we will present two different studies to show how we deal with missing values for the air quality dataset from K-EPA (Data Imputation: Study 1) and the patients with rheumatoid disease activity dataset from KRRD (Data Imputation: Study 2).

4.1 Study 1: Dealing with Environmental Missing Data - Application on K-EPA Data

Incomplete data may arise due to several different reasons including refusal, attrition, measurement errors or simply ignorance about the individual question asked. No matter what the reason is, missing observations is a problem that has to be dealt with in all statistical areas (Allison, 2001). Besides making sure that the missing observations really are missing observations (Schafer and Graham, 2002), assumptions, explicitly or implicitly, about the missing data mechanism are always made. Assuming an ignorable missing data mechanism simplifies the analysis of the missing data as it means that the process causing the missing observations does not have to be explicitly modelled. The concept of ignorable missing data was introduced by Rubin (1976) as "the weakest simple conditions on the process that causes missing data such that it is always appropriate to ignore this process when making inferences about the distribution of the data". For the missing data mechanism to be ignorable, two conditions have to be fulfilled. First, the missing observations have to be missing at random (MAR). Second, the parameters in the missing data process have to be distinct from those in the the data.

The missing data pattern, describing which observations in the data are missing, may further be useful to examine when dealing with incomplete data. A monotone missing

data pattern (MMP) offers, for example, more flexibility in the choice of the missing data method than an arbitrary missing data pattern (AMP) (Little and Rubin, 2002).

In this study, we will present how we are dealing with missing data for the environmental dataset using the missing data imputation technique. The imputation methods used in this work are: multivariate imputation by chained equations (MICE) using random forest (RF_m), k-nearest neighbour (kNN), Bayesian principal component analysis (BPCA), multiple imputation using expectation maximization with bootstrapping (EM with Bootstrapping), predictive mean matching (PMM), and the proposed iterative imputation method (missForest) based on a random forest (Buuren and Groothuis-Oudshoorn, 2010; Shah et al., 2014; Van Buuren et al., 2015; Stekhoven and Bühlmann, 2012; Misztal, 2013; Stacklies et al., 2016). The root mean square error (RMSE) and mean absolute error (MAE) criteria are used to compare the performances of the imputation methods. For the error indicators (RMSE or MAE), the larger the value, the greater the error. The end product is an outline of the best approaches for managing missing data in a data set that is critical for public health in Kuwait.

4.1.1 Missing Data Imputation for the K-EPA Dataset

In environmental research, missing data are often a challenge for statistical modelling. This work addressed some advanced techniques to deal with missing values in a data set measuring air quality using a multiple imputation (MI) approach. MCAR, MAR, and MNAR missing data techniques are applied to the data set. Five missing data levels are considered: 5%, 10%, 20%, 30%, and 40%. The imputation method used in this study is an iterative imputation method, missForest, which is related to the random forest approach. Air quality data sets were gathered from five monitoring stations in Kuwait, aggregated to a daily basis. Logarithm transformation was carried out for all pollutant data, in order to normalise their distributions and to minimize skewness. We found high levels of missing values for NO_2 (18.4%), CO (18.5%), PM_{10} (57.4%), SO_2 (19.0%), and O_3 (18.2%) data. Climatological data (i.e., air temperature, relative humidity, wind direction, and wind speed) were used as control variables for better

estimation. The results show that the MAR technique had the lowest RMSE and MAE. In this chapter, we conclude that MI using the missForest approach has a high level of accuracy in estimating missing values. MissForest had the lowest imputation error (RMSE and MAE) among the other imputation methods and, thus, can be considered to be appropriate for analysing air quality data.

4.1.2 The Review of Missing Imputation

Air quality monitoring is conducted with the aim of protecting public health. Numerous air contaminants have been found to have harmful effects on human health. The air quality in cities varies, due to concentrations of particulate matter up to 10 micrometres in diameter (PM_{10}), NO_2 , O_3 , CO , and SO_2 , from emission sources including vehicle exhaust, manufacturing operations, and chemical facilities, among other sources.

A major challenge in air quality data management is determining how to deal with missing data values. Missing information in data sets occurs for multiple reasons, such as impaired equipment, insufficient sampling frequency, hardware problems, and human error (Norazian et al., 2008). Incomplete data sets affect the applicability of specific analyses, such as receptor modelling, which generally requires a complete data matrix. The occurrence of missing data, no matter how infrequent, can bias findings on the relationships between air contaminants and health outcomes (Junger and de Leon, 2009). Incomplete data matrices may provide outcomes that vary significantly, compared to the results from complete data sets (Forbes et al., 2004).

To gain a more complete data set, researchers must decide whether to discard or impute (i.e., substitute for) missing data. Ignoring missing values is typically not warranted, as valuable information is lost, which may compromise inferential power (Jadhav et al., 2019). Therefore, the most appropriate option is to impute the missing data. Yet, the systematic differences between real and substituted data can also lead to unwanted bias. Therefore, it is vital to determine an optimal approach for estimating missing values. Several problems have been linked with missing data (Hawthorne et al., 2005). These challenges include statistical power reduction, bias as a result of inconsistent

data, difficulties in managing the data during statistical analyses, and low efficiency. The criteria implemented for measures to deal with missing data in time series analysis rely on the missing data replacement mechanism and missing data pattern (Plaia and Bondi, 2006). Such challenges are especially problematic when the missing data exceed 60 percent, where existing methods have significant difficulty in addressing such situations (Farhangfar et al., 2008).

This study focuses on a case study of missing data related to air quality monitoring. The Kuwait environmental public authority (K-EPA) is mandated with the responsibility for measuring air quality. A data set collected from five fixed monitoring stations was associated with missing data, likely caused by multiple reasons. One is that there were a large number of routine maintenance changes in the monitoring sites. Second, simple human error occurred. Third, there were some tagging problems that necessitated the exclusion of some data (e.g. programming or coding issue(s), data structures, backup or archiving process, ..., etc.).

The main purpose of this chapter is to find the best imputation method to estimate the missing values for the monitored pollutants (SO_2 , NO_2 , CO , O_3 , and PM_{10}) from K-EPA datasets using multiple imputation methods (RF_m , kNN, BPCA, EM with Bootstrapping, PMM and missForest).

It is important to describe the factors that may lead to missing data in statistical analyses. The first instance of missing data is missing completely at random (MCAR), whereby the missing data result from either the observer not collecting the necessary information or the reporting of incomplete or false information. The second instance of missing data is missing at random (MAR), whereby the extent of data missing depends on the type of data under observation. MAR is appropriate when the missing data can be partially retrieved, depending on the existence of information related to the variables in the same data set. The third instance is missing not at random (MNAR), whereby the missing data are dependent on the actual values absent for statistical analysis. Among the three types of missing data in statistical analysis, MAR and MNAR are the most common (Graham, 2009). When the type of missing data tends towards MAR, multiple

imputation techniques are more suitable than other techniques, such as listwise deletion (Rubin, 1996).

4.1.3 Missing Data Mechanisms

Rubin (1987) outlined three mechanisms behind missing data in his seminal article: Missing at Random (MAR), Missing Completely at Random (MCAR), and Missing Not at Random (MNAR).

Missing Completely at Random (MCAR)

When missingness on a variable is independent to the variable's values or any other measured variable, the data may be called MCAR (Rubin, 1976; Little and Rubin, 2019; Van Buuren, 2018). In essence, the observed data constitute a random sampling of the entire data set. That is, there are no consistent variations between participants who have missing data and those who have complete data. Because a batch of lab samples was poorly handled, some participants may have missing laboratory values. The missing data diminish the study's analysable population and, as a result, its statistical power, but they do not generate bias: if the data is considered MCAR, the remaining data can be categorized a simple random sampling of the entire data set of interest. The MCAR assumption is considered as a strong and frequently unreasonable assumption.

In MCAR, the chance of missing data values is the same across all instances. The following example describes an instance in which MCAR occurs in statistical analysis:

Suppose that Y is an $n \times p$ matrix which includes all p variables with n cases in the sample. Let the observed values be denoted as (Y_{obs}) , while the missing values are denoted as (Y_{mis}) . The matrix R spots the missing values' locations in Y . The observations of R and Y are denoted as r_{ij} and y_{ij} , respectively. Thus, $r_{ij} = 1$ when y_{ij} is observed, while $r_{ij} = 0$ when y_{ij} is missing. Then, the distribution of R depends upon $Y = (Y_{\text{obs}}, Y_{\text{mis}})$. We can write $\Pr(R|Y_{\text{obs}}, Y_{\text{mis}}, \Psi)$ when the data are said to be assumed as MCAR, if:

$$P(\mathbf{R} \mid \mathbf{Y}) = \Pr(R = 0 | Y_{\text{obs}}, Y_{\text{mis}}, \Psi) = \Pr(R = 0 | \Psi), \quad (4.1)$$

where Ψ consists of the parameters of the missing data in the model. This means that the probability of missing a data value depends only on the estimated parameters in the model.

Missing at Random (MAR)

When missingness on a variable is associated with the observed data but not the unobserved data, the data is termed MAR (Rubin, 1976; Little and Rubin, 2019). If male respondents are less likely to finish a survey on depression gravity than female respondents, a researcher studying depression may come into data that is MAR. In that case, if the likelihood of completing the survey is connected to their sex (which is completely observed) but not to the intensity of their depression, the data can be classified as MAR. Complete case studies of a data set comprising MAR data, which are founded on only data points for which all relevant data is present and no fields are missing, can or cannot result in bias. However, if the entire case analysis is skewed, adequate accounting for known factors (in this case, sex) can give impartial study results.

MAR is, therefore, a less stringent assumption, compared to MCAR; for instance, when selecting a sample from a population based on certain characteristics, the resulting missing data can be categorized as MAR. Statistical software for multiple imputations usually assumes that the data are MAR (Little and Rubin, 2019; Paik and Sacco, 2000). Therefore, the probability of data missing is dependent on the observed data but not the unobserved values:

$$P(\mathbf{R} \mid \mathbf{Y}) = \Pr(R = 0 | Y_{\text{obs}}, Y_{\text{mis}}, \Psi) = \Pr(R = 0 | Y_{\text{obs}}, \psi). \quad (4.2)$$

The K-EPA data are best classified as MAR.

Missing Not at Random (MNAR)

When missingness on a variable is linked to the unobserved data values, the data is considered MNAR (Polit and Beck, 2008; Rubin, 1976; Little and Rubin, 2019). To expand on the preceding scenario, the depression registry may come across MNAR data if people with acute depression have a high likelihood of declining to fill out the depression gravity survey. Complete case scrutiny of a data set encompassing MNAR data, as for MAR data, has the propensity to be biased or not; if it is, this matter cannot be solved in analysis, and estimated effects will be biased.

For MNAR, the chance of data not being available is dependent on reasons unknown to the researcher. For instance, when conducting research, some respondents may decide to withhold information for reasons unknown to the researcher. Due to the nature of MNAR, it is often regarded as a more complex case in statistical analysis. It can be addressed by targeting some of the reasons respondents would choose to withhold information, Y_{mis} , itself. It is represented:

$$P(\mathbf{R} \mid \mathbf{Y}) = \Pr(R = 0 \mid Y_{\text{obs}}, Y_{\text{mis}}, \Psi). \quad (4.3)$$

4.1.4 Ignoring the Missing Data Mechanism

One of the major issues that arise when performing imputations is whether the missing data come from the same distribution as the observed data (Y_{obs}) (Schafer, 1997). As mentioned above, the observed data are made up of Y_{obs} and R with the joint density function $f(Y_{\text{obs}}, R \mid \theta, \Psi)$, which depends on the model estimated parameters θ for Y (Little and Rubin, 2002).

We can estimate θ without knowing Ψ by defining the probability density function of the joint distribution of Y_{obs} and Y_{mis} as $f(Y \mid \theta) \equiv f(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta)$. Therefore, in order to compute the marginal probability density of Y_{obs} , we integrate the missing data as:

$$f(Y_{\text{obs}} \mid \theta) = \int f(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta) dY_{\text{mis}}, \quad (4.4)$$

where the likelihood function of θ , according to Y_{obs} while ignoring the missing data, can be defined as:

$$L_{\text{ign}}(\theta|Y_{\text{obs}}) \propto f(Y_{\text{obs}}|\theta). \quad (4.5)$$

To build a more general model, we include R and specify the joint density distribution of Y and R as:

$$f(Y, R|\theta, \Psi) = f(Y|\theta)f(R|Y, \Psi). \quad (4.6)$$

We can find the distribution of the observed data by integrating Y_{mis} from the joint density using θ and Ψ , defined as:

$$f(Y_{\text{obs}}, R|\theta, \Psi) = \int f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) f(R|Y_{\text{obs}}, Y_{\text{mis}}, \Psi) dY_{\text{mis}}. \quad (4.7)$$

Now, we can rewrite Equation (4.7) as:

$$f(Y_{\text{obs}}, R|\theta, \Psi) = f(R|Y_{\text{obs}}, \Psi) \int f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) dY_{\text{mis}} = f(R|Y_{\text{obs}}, \Psi) f(Y_{\text{obs}}|\theta). \quad (4.8)$$

The missing data mechanism is ignorable for likelihood inference if the following hold

1. MAR: when the missing data pattern is missing at random; and
2. Distinctness: when the joint parameter space of (θ, Ψ) is equal to the product of the parameter space of θ and Ψ (Schafer and Olsen, 1998).

4.1.5 Multiple Imputation (MI)

Some studies investigated the bias and efficiency in data sets with increasing proportions of missing data (e.g. when it exceeds 50% of the total missing values) (Haji-Maghsoudi et al., 2013; Lee and Carlin, 2012; Marshall et al., 2010; McNeish, 2017; Clavel et al., 2014). Researchers have debated the role of listwise deletion when solving for such

missing data. Most research studies have concluded that, although the listwise deletion technique is not commonly used, it is applicable in some instances (Heitjan and Rubin, 1990; King et al., 2001). According to Marshall et al. (2010), multiple imputation is favourable for computing missing data and especially applicable when the missing data rate is above 10% (Newman, 2014). For instance, in a regression model, including a number of variables with a low rate of missing data in the full regression model, when compared to the outcomes of simple bivariate regressions. Therefore, it is critical for analysts to evaluate the total missing rate over all variables as well as the partial missing one for each variable.

One limitation of applying a single imputation approach is that formulas of standard variance applied to filled-in data tend to underestimate the variance of the estimates; therefore, multiple imputation methods have been proposed (Little and Rubin, 2019). The first step in such a method is specifying the single encompassing multivariate approach for all data sets. There are four types of multivariate models of data completion to consider (Schafer and Olsen, 1998): (i) standard models, which impute under multivariate normal distributions; (ii) log-linear models, that have been used traditionally by social scientists in describing the associations among cross-classified data variables; (iii) general location models, which combine the log-linear approach for the variables in the multivariate model of standard regression for the continuous variables; and (iv) a two-level model of linear regression, which is mostly applied to multi-level data. The imputation model should be able to adopt the subsequent analysis and should be able to preserve the interactions of variables, which relates to the central point of the investigation discussed later in this chapter.

A multiple imputation method balances ease of application and the quality of obtained results. The various imputations identify random errors that are appropriate to the process of imputation, making it possible to obtain unbiased estimates in all parameters. No deterministic method of imputation can achieve the same result. The technique also allows for departure from normality assumptions, while providing results that are adequate with low sample sizes or when significant amounts of data are missing.

Some requirements are necessary, in order to attain the desired results of multiple imputation (Allison, 2000). First, there should be random data missing (MAR), which means that there is a dependence on observed variables and not missing observations. Second, the method of generating the values imputed should suit the analysis that subsequently follows. This maintains the associations between variables, which is a focus in the analysis shown later in this chapter. Rubin has given a thorough description of these conditions. A remaining question, however, relates to adopting the most suitable practices for performing the imputations (White et al., 2010). It is essential to have an awareness of the possible prediction problems, in order to reduce or minimize systematic error.

There have been many applications of multiple imputation in health, environmental (Allen and DeGaetano, 2001; Kotsiantis et al., 2006), and industrial (Jagannathan and Wright, 2008; Lakshminarayan et al., 1999) data bases, as well as for survey data (Van Ginkel et al., 2007; Schenker and Taylor, 1996) and data mining approaches, which extract patterns from large data sets through a combination of artificial intelligence and statistical methods, that can be used for database management (Jagannathan and Wright, 2008).

4.1.6 Multiple Imputation Using Random Forest Method

Let us assume that $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p)$ is a $n \times p$ -dimensional data matrix. We propose the use of the random forest technique for imputing missing observations. The random forest algorithm has a built-in routine to handle the values that are missing. This is achieved by weighing the frequency of values with the proximity of bagging modification. Consequently, this builds a large collection of de-correlated trees, and then averages them after the training of an initially imputed mean data set (Breiman, 2001). This approach requires a response variable that is complete and useful for forest training. Instead, we estimate the values of all the missing values directly, by use of a random forest that is trained on the observed data set, where X is the matrix of the complete data. \mathbf{X}_s contains all missing values at entries $\mathbf{i}_{mis}^{(s)} \subseteq \{1, \dots, n\}$. The data set

can be separated into four parts:

1. $\mathbf{y}_{obs}^{(s)}$: the observed values of \mathbf{X}_s .
2. $\mathbf{y}_{mis}^{(s)}$: the missing values of \mathbf{X}_s .
3. $\mathbf{x}_{obs}^{(s)}$: the observations, $\mathbf{i}_{obs}^{(s)} = \{1, \dots, n\} \setminus \mathbf{i}_{mis}^{(s)}$, that belong in the other variables \mathbf{X}_s .
4. $\mathbf{x}_{mis}^{(s)}$: the observations, $\mathbf{i}_{mis}^{(s)}$, that belong in the other variables \mathbf{X}_s .

Note that $\mathbf{x}_{obs}^{(s)}$ and $\mathbf{x}_{mis}^{(s)}$ are not completely observed, as the index $\mathbf{i}_{obs}^{(s)}$ corresponds to the observed values of the variable \mathbf{X}_s .

According to Stekhoven and Bühlmann (2012), the process starts with an initial guess for the missing values in \mathbf{X} using a mean imputation approach or any other imputation method, depending on the data. Then, we sort the predictors $\mathbf{X}_s, s = 1, \dots, p$, ascending or descending, $\mathbf{X}_s, s = 1, \dots, p$, according to the number of missing values. Then, for each variable \mathbf{X}_s , the missing values are imputed by random forest (i.e., the first fitting) with response $\mathbf{y}_{obs}^{(s)}$ and predictors $\mathbf{X}_{obs}^{(s)}$. Next, the missing values $\mathbf{y}_{mis}^{(s)}$ are estimated by applying the trained random forest to $\mathbf{x}_{mis}^{(s)}$. The imputation approach should be repeated until a stopping criterion is reached. Pseudo Algorithm 1 shows a representation of the missForest method (see Algorithm 1).

The stopping criterion (γ) is met when the difference between the last imputed data matrix and the previous one increases for the first time, with respect to both variable types. Here, the difference for the set of continuous variables \mathbf{N} is defined as:

$$\Delta_N = \frac{\sum_{j \in \mathbf{N}} (\mathbf{X}_{new}^{imp} - \mathbf{X}_{old}^{imp})^2}{\sum_{j \in \mathbf{N}} (\mathbf{X}_{new}^{imp})^2}, \quad (4.9)$$

and that for the set of categorical variables \mathbf{F} as:

$$\Delta_F = \frac{\sum_{j \in \mathbf{F}} \sum_{i=1}^n \mathbf{I}_{\mathbf{X}_{new}^{imp} \neq \mathbf{X}_{old}^{imp}}}{\#NA}. \quad (4.10)$$

Algorithm 1 Impute missing values with random forest, Stekhoven and Bühlmann (2012).

Require: \mathbf{X} is an $n \times p$ matrix. Set up the stopping criterion (γ)

- 1: set up initial guess for missing values;
 - 2: \mathbf{k} is the vector of sorted indices of columns in \mathbf{X} w.r.t. increasing the amount of missing values;
 - 3: **while** not γ **do**
 - 4: \mathbf{X}_{old}^{imp} stores the previously imputed matrix;
 - 5: **for** s in \mathbf{k} **do**
 - 6: Fit a random forest: $\mathbf{y}_{obs}^{(s)} \sim \mathbf{x}_{obs}^{(s)}$;
 - 7: Predict $\mathbf{y}_{mis}^{(s)}$ using $\mathbf{x}_{mis}^{(s)}$;
 - 8: \mathbf{X}_{new}^{imp} updates the imputed matrix using predicted $\mathbf{y}_{mis}^{(s)}$;
 - 9: **end for**
 - 10: update γ
 - 11: **end while**
 - 12: **return** the imputed matrix \mathbf{X}^{imp}
-

Let \mathbf{X} be an $n \times p$ matrix; set the stopping criterion (γ); set the initial guess for missing values. $\mathbf{k} \leftarrow$ vector of sorted indices of columns in \mathbf{X} w.r.t. increasing amount of missing values. $\mathbf{X}_{old}^{imp} \leftarrow$ stores the previously imputed matrix. Fit a random forest: $\mathbf{y}_{obs}^{(s)} \sim \mathbf{x}_{obs}^{(s)}$. Predict $\mathbf{y}_{mis}^{(s)}$ using $\mathbf{x}_{mis}^{(s)}$; $\mathbf{X}_{new}^{imp} \leftarrow$ update the imputed matrix using the predicted $\mathbf{y}_{mis}^{(s)}$. Update γ and the imputed matrix \mathbf{X}^{imp} , where #NA is the number of missing values in the categorical variables \mathbf{F} .

After imputing the missing values, the performance is assessed using the normalised root mean squared error (Oba et al., 2003) for the continuous variables, defined by:

$$\text{NRMSE} = \sqrt{\frac{\text{mean} \left((\mathbf{X}^{\text{true}} - \mathbf{X}^{\text{imp}})^2 \right)}{\text{var} (\mathbf{X}^{\text{true}})}}, \quad (4.11)$$

where \mathbf{X}^{true} and \mathbf{X}^{imp} are the complete data matrix and the imputed data matrix, respectively. In this study, all predictors are classified as continuous observations. The mean and variance are used as a short notation for empirical mean and variance computed over the missing values only.

When an RF_m is fitted to the part that is observed on a variable, we use the out-of-bag (OOB) estimate of an error for the variable. When we meet the stopping criterion (γ), we average it over the variable set of that type, in order to obtain an approximation of the

actual errors of imputation. We assess the performance of this estimate by comparing the absolute difference between the OOB imputation error estimate in all simulation runs and the true imputation error.

4.1.7 Process of Multiple Imputations (MI) Using Rubin's Rules

For our data sets, we followed Rubin's rules (Little and Rubin, 2019) for handling missing data. The process of multiple imputations (MIs) was conducted separately for each monitoring station (see figure 4.1). The first step in multiple imputation is to create values ("imputes" or " m_i "), with 10 iterations for each " m_i " to be substituted for the missing data. In order to create imputed values, we need to identify a model (say, a linear regression) that allows us to create imputes based on other variables in the data set (predictor variables). As we need to do this multiple times, in order to produce multiple-imputed data sets, we identify a set of regression lines which are similar to each other.

Figure 4.1 shows the process for the K-EPA data sets, to process and estimate missing values using imputation methods. There were five data sets (1-5), relating to FAH, JAH, MAN, RUM, and ASA, respectively. Each data set should contain 2192 daily observations for each variable; however, due to missing values, they were all less than 2192.

The power of MI lies in its multiple imputations being able to be performed for each variable in the data set. While every single imputation is ambiguous or imprecise, the combination of the computed imputations takes the uncertainty of each imputation into consideration. According to King et al. (2001) and Newman (2014), MAR or MCAR pooled estimated parameters are less biased and the associated standard errors are corrected appropriately.

The implementation of an MI technique requires three steps: First, it imputes several values for the same observation, using at least two different imputed values ($m \geq 2$) for each MI approach. Then, the second step takes each individual method, m , and analyses it using standard complete data. Finally, the completed data sets are pooled

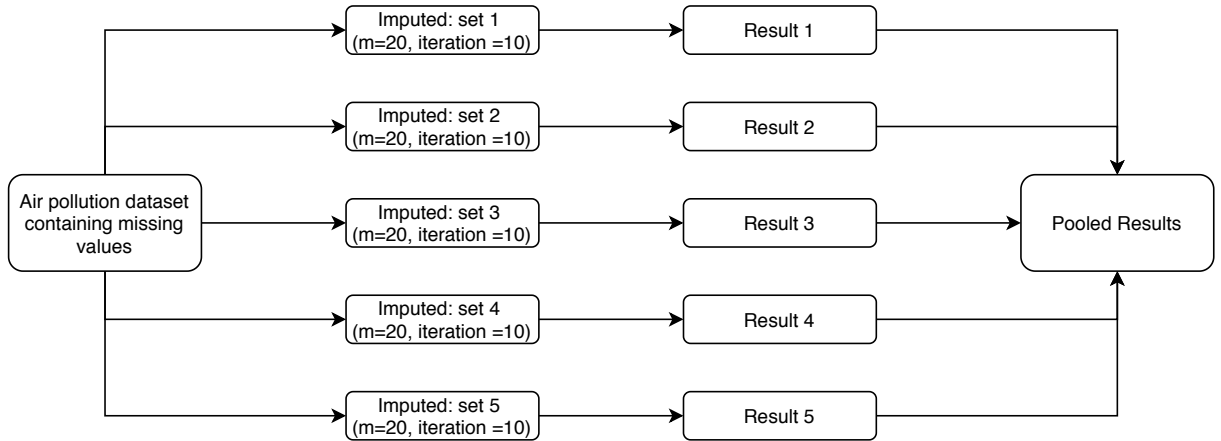


Figure 4.1: The steps of implementing multiple imputations for PM_{10} , SO_2 , O_3 , CO , and NO_2 during 2012 to 2017 with 20 imputed datasets ($m=20$) according to site location, in the State of Kuwait.

by integrating the m analyses, in order to generate overall estimates and standard errors. This can be done by calculating the mean over the m repeated analyses. Pooling data from several m allows multiple imputations to ensure higher accuracy (Ghazali et al., 2020). Figure 4.1 shows how we treated the K-EPA data sets with multiple imputation, where $m = 20$.

4.1.8 Data Sets from Kuwait EPA

We have utilised a real-time air quality monitoring data set collected for 5 locations in Kuwait from the Kuwait Environmental Public Authority (K-EPA), in order to evaluate and assess the performance of various imputation methods to estimate missing values in the data set. The data set contained air quality, time, and meteorological data.

1. Air quality data: The air pollutant variables in the air quality data were NO_2 , CO , PM_{10} , SO_2 , and O_3 ;
2. Meteorological data: The meteorological parameters included temperature, humidity, wind direction, and wind speed.

We compiled pollutant data from the Environmental Public Authority of Kuwait

(K-EPA). The data were gathered from five environmental monitoring stations from 1 January 2013 to 31 December 2017. Based on the daily data, both the 24 hours aggregation for SO_2 , NO_2 , and PM_{10} , and the 8 hours aggregation for CO and O_3 at each station were calculated. All pollutants were measured using the micrograms per cubic metre ($\mu g/m^3$). According to US Environmental Protection Agency (2015), if less than 75% of data are present (i.e. less than 6 hours), the average is considered missing. We used the Air Quality Index (AQI), as generated by Al-Shayji et al. (2008).

The AQI was developed, for Kuwait, based on the United States Environmental Protection Agency (USEPA) recommendations. As mentioned in section 1.3.5 on page 11, the AQI calculation was explained and performed using equation 1.1 on page 13 and table 1.2 on page 13.

Using the data obtained from K-EPA, we conducted an in-depth comparative analysis of the different imputation methods. Missing data were entered into each data set, assuming a general missing data pattern and three mechanisms of missing data: MCAR, MAR, and MNAR. Under the MCAR assumption, missing values were randomly applied to each data set. Under the MAR assumption, the probability of information being missing depended on class attribute. Under the MNAR assumption, the largest or smallest values of X_s were removed. The objective of the study was to derive a comparison of six different imputation methods for MNAR, MAR, and MCAR, concerning missing data. We simulated the rates of missing data by varying the proportions by 5%, 10%, 20%, 30%, and 40%. We used "ampute" function in "MICE" package in R to generate multivariate missing data under a MCAR, MAR or MNAR missing data mechanism.

4.1.9 Missing Imputation Evaluation Criteria

To determine the best imputation method, three model performance measures were considered (Bennett et al., 2013): root mean square error (RMSE), mean absolute error (MAE), and correlation coefficient (R), which are calculated as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (4.12)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (4.13)$$

where y_i and \hat{y}_i are the i th observations for the comparison value and the imputed data sets, respectively. The error was measured based on the difference between the estimated value and the observed values. For RMSE and MAE measures, if the value obtained is small, then the estimation method is better.

4.1.10 R Packages Used for Imputation Process

Five well-known imputation packages accessible in R were applied. The first R package used here was VIM (<https://cran.r-project.org/web/packages/VIM/VIM.pdf>), which is associated with kNN imputation methods and robust model-based imputation for numerical, semi-continuous, categorical, or ordered variables (Kowarik and Templ, 2016). The second R package was MICE (<https://cran.r-project.org/web/packages/mice/mice.pdf>) which stands for Multivariate Imputation via Chained Equations (Royston, 2004). MICE is specialised to deal with missing values of MAR or MNAR types (Buuren and Groothuis-Oudshoorn, 2010). MICE can deal with different types of variables using different imputation methods, such as predictive mean matching for numeric variables, logistic regression for binary variables, Bayesian polytomous regression for factor variables, and a proportional odds model for ordered variables (Buuren and Groothuis-Oudshoorn, 2010; Horton and Lipsitz, 2001). The third package was missForest (<https://cran.r-project.org/web/packages/missForest/missForest.pdf>). MissForest deals with non-parametric imputation (Stekhoven and Bühlmann, 2012). MissForest enables the imputation of the predictors by using regression trees of resampling under the prediction of missing values (Liao et al., 2014). MissForest has good computational efficiency and can work well with high-dimensional data (Stekhoven and Bühlmann, 2012). The fourth package was Amelia

(<https://cran.r-project.org/web/packages/Amelia/Amelia.pdf>), which enables imputation by maximizing the level of expectation with a bootstrapping algorithm. The Amelia package has also been recommended under a larger number of variables with high-dimensional data. The package also provides improved imputation models by adding Bayesian priors on individual cell values (Honaker et al., 2011). The final package used was missCompare (<https://cran.r-project.org/web/packages/missCompare/missCompare.pdf>). The missCompare package provides several diagnostic measurements to compare between all imputation methods, using RMSE, MAE, and other imputation performance criteria.

4.1.11 Statistical Results

Based on results for the real-time ambient air quality and meteorological data from the monitoring stations in K-EPA, we inferred real-time and fine-grained ambient air quality information using means and standard deviations. The distribution analysis was conducted using the skewness and kurtosis with information of the quartiles (e.g., 25th and 75th quartiles, median, and IQR), where the correlation between the predictors was assessed by the Pearson correlation coefficient. The rate of missing values is presented for each monitoring station using the percentage of total number of missing values among the predictors.

Table 4.1 shows the average air pollutant concentrations. The overall mean and SD for PM_{10} , CO , NO_2 , O_3 , and SO_2 were 0.23 ± 1.07 , 0.91 ± 0.90 , 0.04 ± 0.02 , 0.02 ± 0.01 , and 0.01 ± 0.01 , respectively. The missing value rates were 52.16%, 19.37%, 22.35%, 22.40%, and 22.93% from all ($N = 9,006$), respectively. Figures B.1 and B.2 from Appendix B.1 show the missing data distribution, based on year and monitoring site. Missing value patterns for air quality measurements from 2012 to 2017 was explained in figure B.3.

All pollutant distributions were positively skewed and we corrected the skewness by applying log transformations (Alsaber et al., 2020). Figure B.4 in the Appendix B.1 shows the distribution performance after we applied logarithmic transformations to

PM_{10} , SO_2 , O_3 , CO , and NO_2 .

Table 4.2 shows the Pearson correlation analysis of various air pollutants and meteorological parameters. The strongest positive correlation was found between NO_2 and SO_2 . This was expected, due to their common emission sources (e.g., road traffic). NO_2 had a weak association with PM_{10} , whereas O_3 had a highly negative association with NO_2 . All meteorological parameters (temperature, humidity, wind speed, and wind direction) showed a negative association with NO_2 .

We performed time series plots for each pollutant and for meteorological parameters (e.g. temperature, wind speed, relative humidity) for each monitoring station to better understand the patterns of the missing data among all observations (see Figures 4.2, 4.3 and 4.4). We concluded that the missing data pattern can be classified as missing at random (MAR) or missing not at random (MNAR), especially for the large missing gaps (see Appendix B, Figures B.1–B.3). Figure B.3 from Appendix B shows missing observation ratios for each pollutant. From Figure B.3, we can conclude that PM_{10} has the highest missing observation rate among the pollutants (see Appendix B Figure B.3-left panel). The right side of the Figure B.3 from Appendix B shows the missing value pattern for each pollutant.

Table 4.1: Distribution of Kuwait ambient air pollution exposure during 2012–2017. The total daily observations for ASA are $N = 1,779$; for FAH, $N = 1,820$; for JAH, $N = 1,819$; for MAN, $N = 1,777$; and, for RUM, $N = 1,811$. 25th is the lower quartile (25th percentile), 75th is upper quartile (75th percentile). SD: standard deviation.

Air Pollutant	ASA	FAH	JAH	MAN	RUM	All (N = 9006)
<i>PM₁₀</i>						
min	0.017	0.004	0.005	0.008	0.019	0.004
25th	0.099	0.076	0.073	0.099	0.121	0.088
median	0.154	0.109	0.107	0.142	0.211	0.140
75th	0.262	0.163	0.180	0.218	0.273	0.232
max	3.248	5.500	1.714	7.216	2.538	7.216
mean (sd)	0.26 ± 0.32	0.17 ± 0.28	0.17 ± 0.20	0.32 ± 2.38	0.25 ± 0.23	0.23 ± 1.07
%Missing	%53.16	%50.43	%53.12	%53.62	%50.48	%52.16
<i>CO</i>						
min	0.050	0.078	0.015	0.048	0.015	0.015
25th	0.597	0.981	0.107	0.719	0.743	0.562
median	0.720	1.265	0.235	0.922	0.971	0.860
75th	0.945	1.567	0.471	1.172	1.241	1.198
max	2.661	3.789	5.956	4.483	68.980	68.980
mean (sd)	0.80 ± 0.32	1.30 ± 0.47	0.36 ± 0.41	0.98 ± 0.41	1.08 ± 1.68	0.91 ± 0.90
%Missing	%21.57	%17.30	%20.57	%19.57	%17.84	%19.37
<i>NO₂</i>						
min	0.001	0.005	0.004	0.001	0.000	0.000
25th	0.028	0.032	0.014	0.018	0.018	0.020
median	0.038	0.045	0.019	0.029	0.026	0.030
75th	0.052	0.066	0.026	0.046	0.039	0.046
max	0.361	0.182	0.095	0.194	0.183	0.361
mean (sd)	0.04 ± 0.02	0.05 ± 0.03	0.02 ± 0.01	0.03 ± 0.02	0.03 ± 0.02	0.04 ± 0.02
%Missing	%20.89	%17.48	%20.89	%34.87	%17.61	%22.35
<i>O₃</i>						
min	0.001	0.002	0.001	0.003	0.001	0.001
25th	0.014	0.012	0.019	0.017	0.015	0.015
median	0.021	0.018	0.025	0.022	0.023	0.022
75th	0.029	0.024	0.033	0.029	0.031	0.029
max	0.073	0.076	0.062	0.065	0.075	0.076
mean (sd)	0.02 ± 0.01	0.02 ± 0.01	0.03 ± 0.01	0.02 ± 0.01	0.02 ± 0.01	0.02 ± 0.01
%Missing	%20.35	%18.48	%20.98	%34.55	%17.66	%22.40
<i>SO₂</i>						
min	0.000	0.000	0.000	0.001	0.001	0.000
25th	0.006	0.005	0.002	0.003	0.005	0.004
median	0.008	0.009	0.003	0.004	0.007	0.006
75th	0.011	0.019	0.005	0.005	0.011	0.010
max	0.038	0.152	0.049	0.058	0.056	0.152
mean (sd)	0.01 ± 0.00	0.02 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.01	0.01 ± 0.01
%Missing	%20.53	%17.39	%22.80	%36.19	%17.75	%22.93

Table 4.2: Correlation analysis between weather climatology and air-pollution components SO_2 , NO_2 , O_3 , CO , and PM_{10} .

	NO_2	O_3	SO_2	CO	PM_{10}	Temp.	RH	WS
NO_2								
O_3	-0.35 **							
SO_2	0.40 **	-0.09 **						
CO	0.35 **	-0.26 **	0.22 **					
PM_{10}	-0.06 **	0.05 *	-0.03 *	-0.03				
Temp.	-0.09 **	0.45 **	-0.06 **	-0.14 **	0.05 *			
RH	-0.02	-0.25 **	-0.08 **	0.29 **	-0.03	-0.61 **		
WS	-0.20 **	0.30 **	0.13 **	-0.22 **	0.10 **	0.24 **	-0.32 **	
WD	-0.25 **	0.13 **	-0.15 **	-0.27 **	0.06 **	0.14 **	-0.28 **	0.31 ***

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Terms: RH: Relative humidity, Temp.: Temperature in Celsius, WS: Wind speed, WD: Wind Direction.

Table 4.3 shows a comparison of missing rates for each monitored pollutant between monitoring stations. There were significant differences among the stations in producing missing values using an ANOVA test (similar to what we have done before in 2.4 on page 28), where all p -values were less than 0.05, except for PM_{10} as $p > 0.05$. PM_{10} was excluded from all imputation calculations, due to a missing rate that exceeded 50% (Zakaria and Noor, 2018; Bertsimas et al., 2017).

Table 4.3: Comparing the differences in Missing data by site using ANOVA test. From the results we conclude that all monitoring fixed stations are different in missing values amount for each pollutant except PM_{10} .

	ASA N = 2192	FAH N = 2192	JAH N = 2192	MAN N = 2192	RUM N = 2192	p -Value
NO_2	454 (20.7%)	379 (17.3%)	454 (20.7%)	761 (34.7%)	382 (17.4%)	<0.001
O_3	442 (20.2%)	401 (18.3%)	456 (20.8%)	754 (34.4%)	383 (17.5%)	<0.001
SO_2	446 (20.3%)	377 (17.2%)	496 (22.6%)	790 (36.0%)	385 (17.6%)	<0.001
CO	469 (21.4%)	375 (17.1%)	447 (20.4%)	425 (19.4%)	387 (17.7%)	0.001
PM_{10}	1163 (53.1%)	1103 (50.3%)	1162 (53.0%)	1173 (53.5%)	1104 (50.4%)	0.069

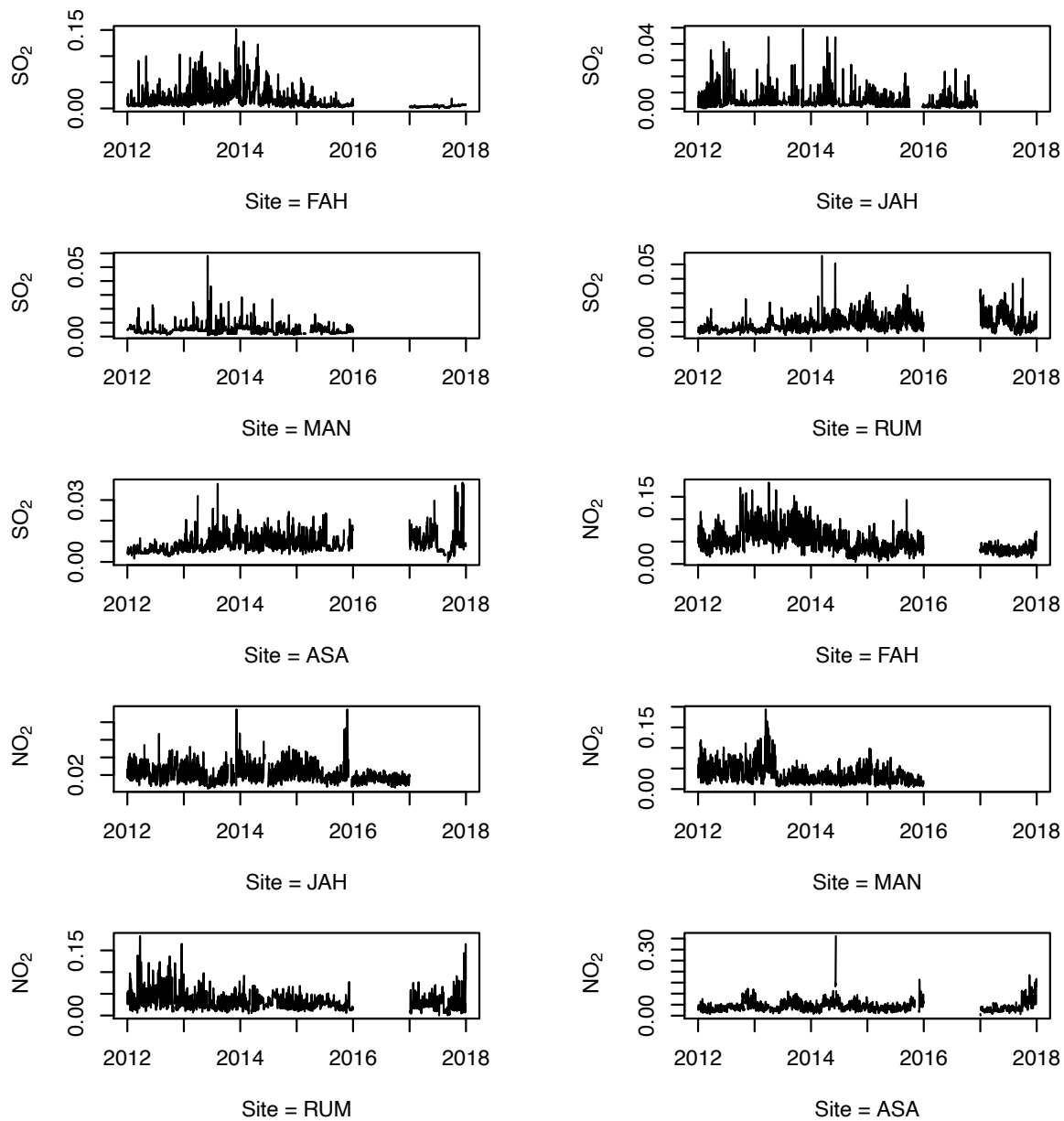


Figure 4.2: Time series of air quality monitoring for SO_2 and NO_2 from 2012 to 2017, with missing values from five different locations (stations) in the State of Kuwait.

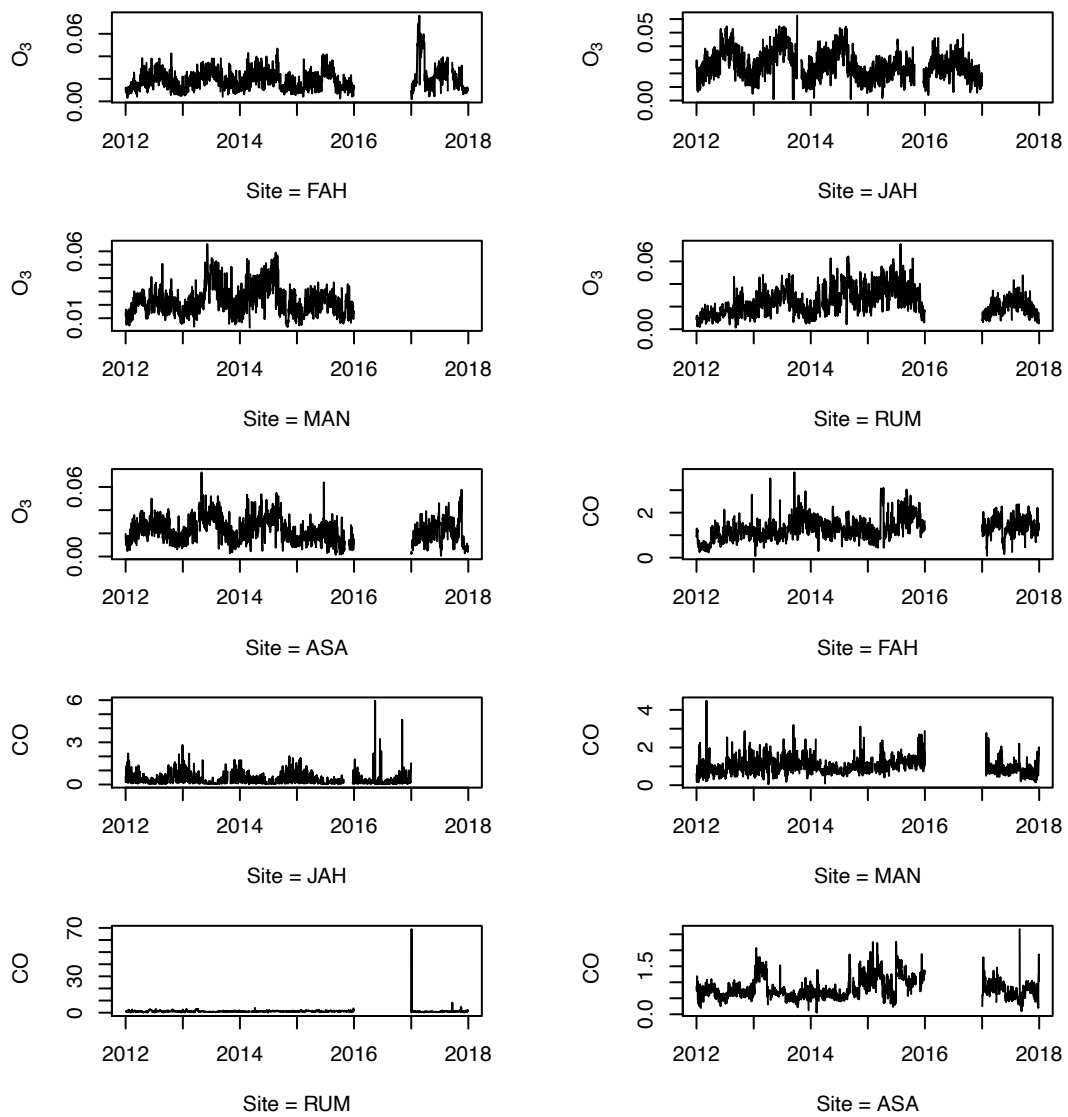


Figure 4.3: Time series of air quality monitoring for O_3 and CO from 2012 to 2017, with missing values from five different locations (stations) in the State of Kuwait.

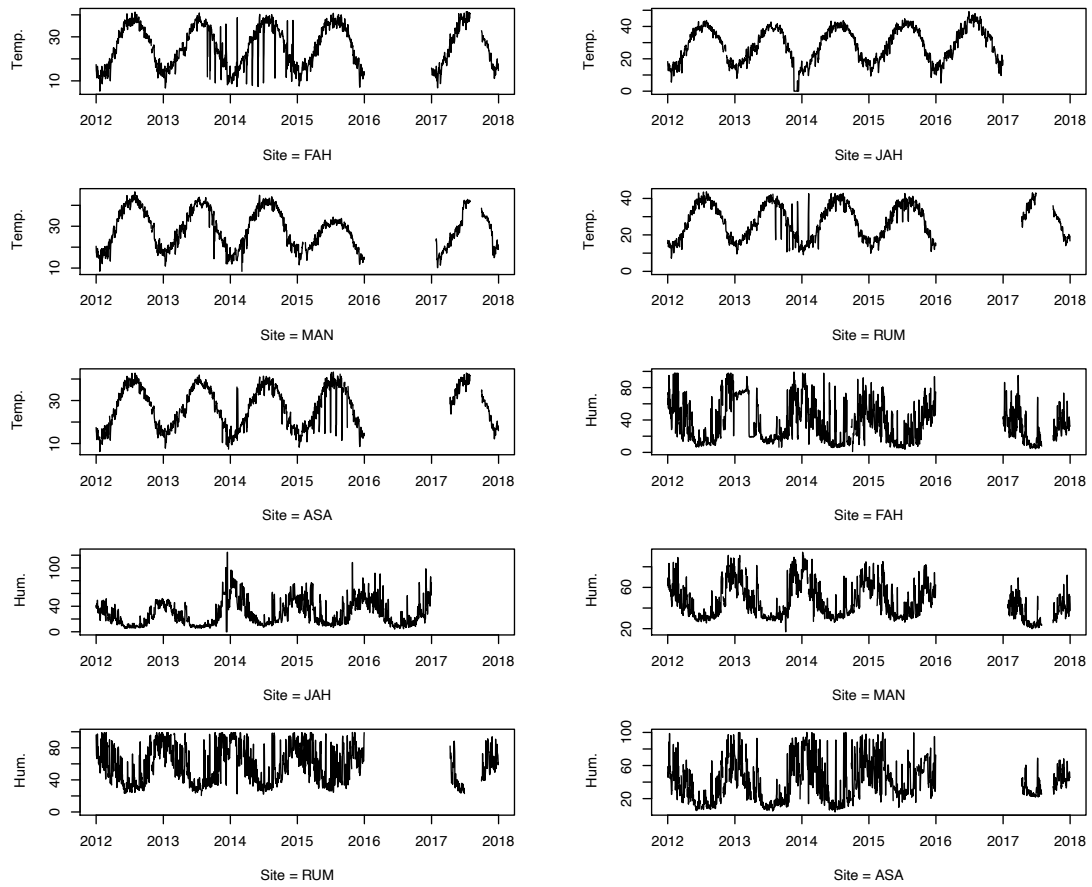


Figure 4.4: Time series of weather climatology (temperature and relative humidity) from 2012 to 2017, with missing values from five different locations (stations) in the State of Kuwait.

4.1.12 Air Quality Missing Data Patterns

As shown in Table 4.4 and Figure B.5 from Appendix B, the RMSE ranged between 1.029 to 2.110 among all methods (EM, PMM, RF_m , missForest, BPCA and kNN) based on all missing mechanisms (MAR, MCAR and MNAR) among all missingness levels (5%, 10%, 20%, 30%, and 40%). The missForest approach performs better among the other imputation methods in all missing mechanisms (MAR, MCAR and MNAR) for all missingness levels (5%, 10%, 20%, 30%, and 40%). This result was consistent with previous studies (Valdiviezo and Van Aelst, 2015; Junger and De Leon, 2015). As seen in Table 4.4 and

appendix Figure B.5, the best imputation method for estimating the simulated missing data was the missForest method. The missForest method had the smallest values of MAE and RMSE for all parameters and percentages of simulated missing data rates, and this finding was consistent with the study of Norazian et al. (2008), where MTB was the best imputation method for filling the missing data, as it was able to obtain the smallest error for all percentages of missing data, in agreement with Kokla et al. (2019); Tang and Ishwaran (2017); Ishak et al. (2017); Valdiviezo and Van Aelst (2015); Shah et al. (2014); Stekhoven and Bühlmann (2012). The second-best imputation method for estimating the simulated missing data was the k-nearest neighbor (kNN) method. The k-nearest neighbor (kNN) was reported as the best imputation approach to fill and estimate the air pollution data by Zakaria and Noor (2018). The worst-performing methods were multiple imputation using additive regression, bootstrapping, and predictive mean matching (PMM) methods. This was also consistent with the study reported by Zakaria and Noor (2018).

Table 4.4: RMSE comparison between the indexed original values and the imputed values using missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR) missingness patterns.

5% Missingness Rate						
Method	RMSE			MAE		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
EM	1.430	1.405	1.536	1.145	1.120	1.238
PMM	1.408	1.430	1.529	1.129	1.140	1.225
RF_m	1.413	1.412	1.547	1.128	1.126	1.242
missForest	1.031	1.035	1.270	0.821	0.823	1.036
BPCA	2.110	1.199	1.568	1.686	0.953	1.251
kNN	1.064	1.065	1.288	0.850	0.846	1.047
10% missingness rate						
Method	RMSE			MAE		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
EM	1.408	1.431	1.517	1.125	1.140	1.218
PMM	1.414	1.415	1.527	1.125	1.131	1.229
RF_m	1.414	1.416	1.529	1.129	1.133	1.231
missForest	1.035	1.028	1.260	0.829	0.820	1.025
BPCA	1.816	1.792	1.813	1.456	1.431	1.449
kNN	1.063	1.064	1.282	0.853	0.846	1.041
20% missingness rate						
Method	RMSE			MAE		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
EM	1.415	1.410	1.523	1.129	1.124	1.225
PMM	1.418	1.417	1.528	1.129	1.131	1.226
RF_m	1.413	1.408	1.532	1.128	1.124	1.228
missForest	1.029	1.038	1.253	0.819	0.827	1.019
BPCA	1.653	1.548	1.856	1.319	1.233	1.478
kNN	1.062	1.065	1.270	0.847	0.850	1.032
30% missingness rate						
Method	RMSE			MAE		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
EM	1.405	1.410	1.531	1.124	1.127	1.232
PMM	1.418	1.419	1.527	1.131	1.132	1.229
RF_m	1.419	1.419	1.521	1.136	1.134	1.224
missForest	1.034	1.033	1.255	0.825	0.823	1.023
BPCA	1.891	1.622	2.060	1.506	1.293	1.645
kNN	1.065	1.064	1.276	0.850	0.848	1.036
40% missingness rate						
Method	RMSE			MAE		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
EM	1.401	1.411	1.518	1.119	1.127	1.222
PMM	1.411	1.399	1.520	1.126	1.116	1.222
RF_m	1.412	1.419	1.534	1.124	1.133	1.234
missForest	1.032	1.035	1.259	0.823	0.827	1.027
BPCA	1.564	1.264	1.789	1.250	1.007	1.428
kNN	1.062	1.067	1.279	0.847	0.852	1.042

From Table 4.3, we can conclude that the missing rates are different among the selected air monitoring stations for each pollutant except PM_{10} that shows similarities in missing rates among the monitoring stations.

The results of the missing data imputation approach were diagnosed using convergent plots for the mean and standard deviation of the multiple imputation data sets using missForest (see Appendix B Figures B.6 and B.7). For convergence, the different streams should not show any definite trends; we did not observe any obvious trends in these data. In addition, Figure B.8 shows kernel density estimates for the marginal distributions of the observed data (blue line) and the $m = 20$ densities per variable calculated from the imputed data (red lines). This indicates stability after 10 iterations.

We imputed the missing information into the original data sets to assess if the imputed data are consistent with the existing data. Figures 4.5 and 4.6 showed how the imputed data sets fit with the actual information in each station. We can see from the figures that large gaps of missing data are filled in the same pattern of the historical values for all pollutants and meteorological parameters which gives a good indication of the suitability of using missForest to estimate missing air pollutant data.

4.1.13 Study 1 - Discussion and Conclusion

In Kuwait, the Environmental Public Authority (K-EPA) is responsible for monitoring the air quality status. The data of air quality obtained from the five stations used in this study usually contain missing data, which can cause bias due to systematic errors between the observed and unobserved values (Alsaber et al., 2020). Therefore, it is vital to determine the optimal approach for estimating the missing values, in order to guarantee that the analysed data are of high quality. Incomplete data matrices may provide outcomes that vary significantly, compared to the results expected from a data set that is complete (Forbes et al., 2004). The primary purpose of any data analysis is to make valid and reasonable inferences on a particular population under study. A researcher is expected to respond to the missing data problem in a way that aligns with the population of interest.

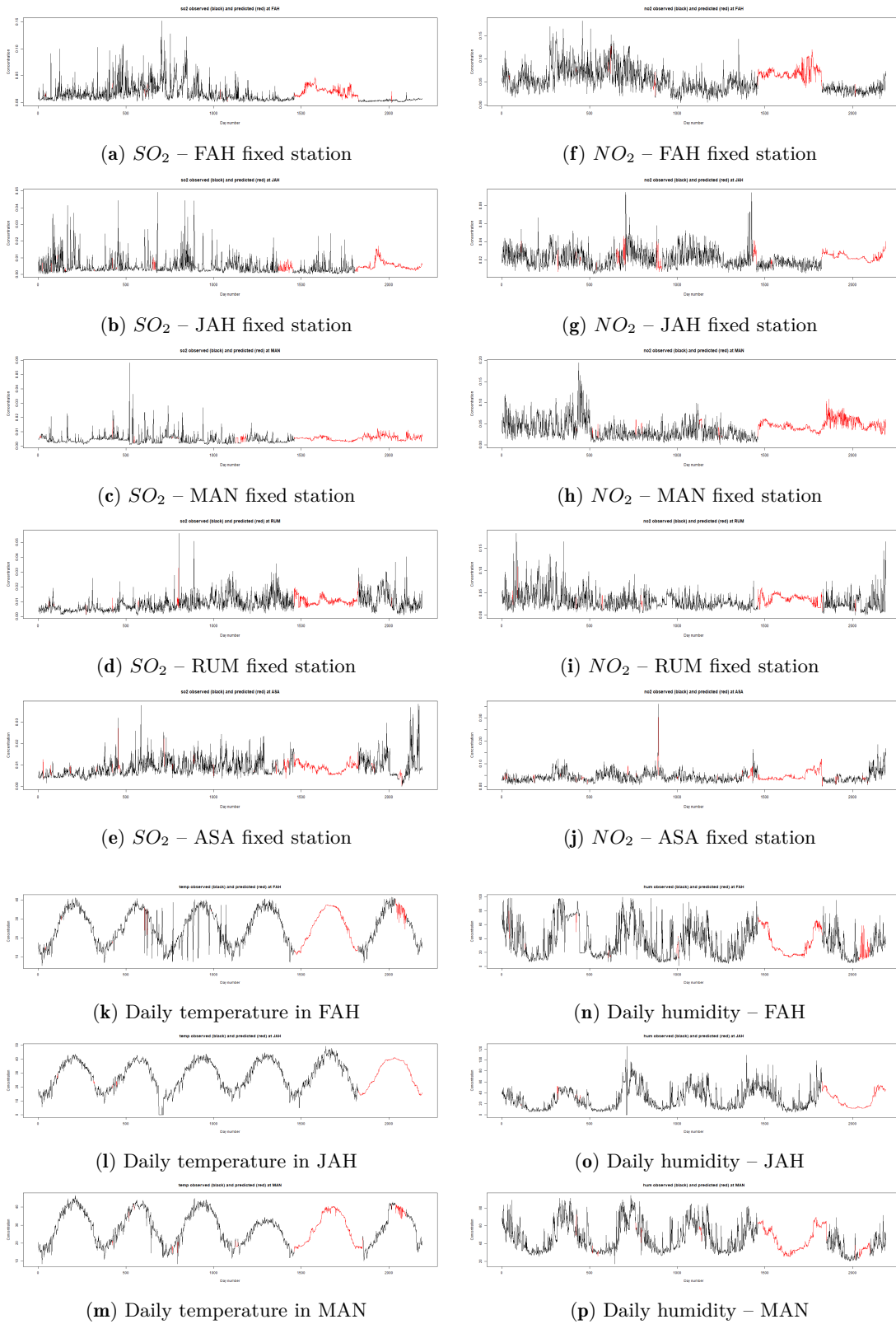


Figure 4.5: Daily concentration for SO_2 , NO_2 , weather temperature and relative humidity after estimating missing values using missForest approach from 2012-2017

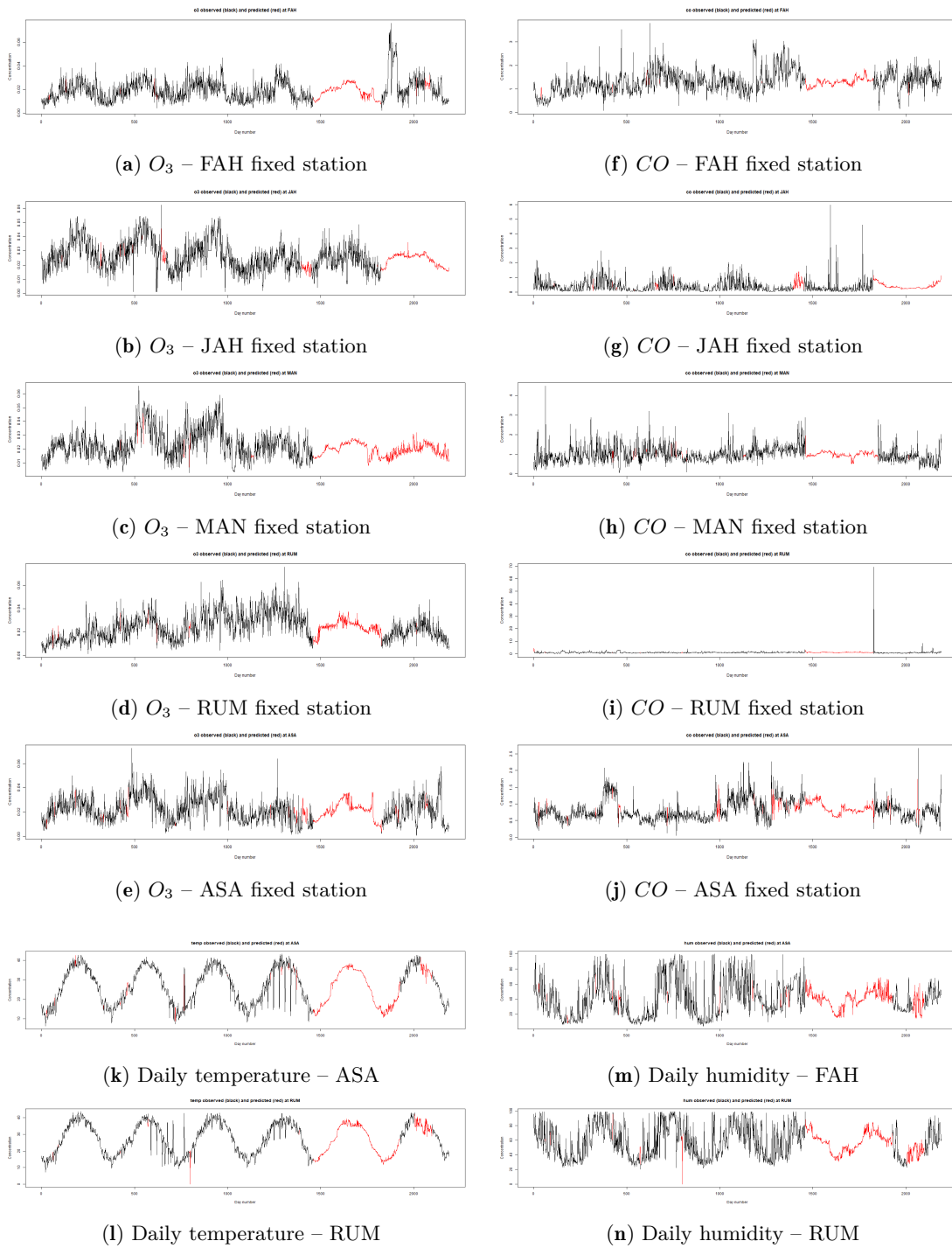


Figure 4.6: Daily concentration for O_3 , CO , weather temperature and relative humidity after estimating missing values using missForest approach from 2012-2017

The main contribution of this work was to find the most appropriate method to fill in missing observations in an air pollution data set from Kuwait. Single and multiple imputation methods were adopted and their performances were compared using the RMSE and MAE metrics. To estimate missing data for SO_2 , NO_2 , PM_{10} , CO , and O_3 in the K-EPA database, we applied artificially introduced missing values ranging from 10% to 40% of the data set. We showed that missForest could successfully handle the missing values, particularly in data sets including different types of environmental variables.

However, this computation method also had limitations. It requires proficiency in R programming, being demanding in comparison to the kNN or PMM methods. There is also a possible connection between the pollutant values and the missing variables. Therefore, these results are not applicable in cases where the missing data are due to non-random reasons. It is evident that some of the observed air pollutant records contained erroneous information. When we ignore this factor during the examination, the results obtained tend to be misleading.

Missing data are always lost, in their entirety and forever, but a proper imputation scheme can help to remedy the situation as much as possible. The method that performs best in each situation, in terms of the assessments, is identified in this work. For this study, missForest gives the most accurate results in estimating the missing values through the multi-dimensional dataset (the datasets that came from five fixed monitoring stations). The missForest method enables imputation on virtually any kind of data. In particular, it can deal with multivariate information comprised of continuous and categorical factors at the same time. This method does not require parameter tuning, nor does it require assumptions about the distribution of the information. Finally, missForest had the least imputation error for each frequency of missingness rates (5%, 10%, 20%, 30%, and 40%), and it had the smallest prediction error difference when models used imputed values.

4.2 Study 2: Dealing with Clinical Missing Data - Application on KRRD

Missing data in clinical epidemiological research violate the intention-to-treat principle, reduce the power of statistical analysis, and can introduce bias if the cause of missing data is related to a patients's response to treatment. Multiple imputation (MI) provides a solution to predict the values of missing data. The main objective of this section is to estimate and impute missing values in patient records. The data from Kuwait Registry for Rheumatic Diseases (KRRD) was used to deal with missing values among the patients' records. A number of methods were implemented to deal with missing data, however choosing the best imputation method was judged by the lowest root mean square error (RMSE). Among 1,735 rheumatoid arthritis (RA) patients, we found the percentage of missing values vary from 5% to 65.5% of the total observations. The results show that the sequential random forest method can estimate these missing values with a high level of accuracy. The RMSE varied between 2.5 and 5.0. MissForest had the lowest imputation error for both continuous and categorical variables under each missing data rate (10%, 20%, and 30%) and had the smallest prediction error difference when the models used the imputed laboratory values.

In much clinical research, missing values or experimental values remain a problem in correctly analysing results and obtaining accurate outcomes. These missing values often lead to misinterpretation and biased results, which could ultimately affect the overall conclusion of an investigation (Sartori et al., 2005; Branden and Verboven, 2009; Alsaber et al., 2021b). The application of statistical analyses in experiments with missing values poses serious problems, as the missing values are often automatically ignored by the statistical algorithms. The results obtained by the investigator in such experiments may be non-significant or even meaningless (Kang, 2013; Stavseth et al., 2019). Missing data is a common problem for all kind of research data, especially in clinical trials. It always becomes problematic when sample collection was not performed in random order or was obtained using an improper methodology (Junninen et al., 2004). Certain factors are

responsible for missing values in the data of a study: (i) the data are not captured due to some unknown reason, such as error in recording the data from an electronic detector/data recorder or manual recording by technical medical staff; (ii) data are missing due to a known reason, such as critical medical conditions; or (iii) data are not recorded as they are unrelated to the patient's clinical medical condition (Kang, 2013). However, the biased and misleading information obtained when values are missing can be managed by the application of imputation methods.

4.2.1 Missing Imputation - Rubin's Approach

Imputation involves the substitution of missing values with known variables. This type of approach is widely used, as it produces complete data. However, the decision regarding the imputed value cannot be unbiased (e.g. multiple imputation for missing data makes it possible for the researcher to obtain approximately unbiased estimates of all the parameters from the random error. The researcher cannot achieve this result from deterministic imputation, which the multiple imputation for missing data can do), as it could lead to an overestimation of confidence in the outcome. To overcome this problem, Rubin suggested the theory of multiple imputation, in which missing values are imputed using the appropriate model a few times (generally three to five times) and a standard method is applied for the analysis (Higgins et al., 2008; Little and Rubin, 2019; Rubin, 1987; Alsaber et al., 2021b). The imputation method provides more accurate results, but problems with the application of imputation include: (1) maximum use of the available data to reduce the error for univariate data and preserve covariance in multivariate data sets; and (2) reporting the variance estimates of uncertainty caused due to the imputed value (Rubin, 1987). Several parametric and non-parametric techniques have been employed to deal with missing values. Parametric methods depend on the assumed method, whereas non-parametric methods require a high number of observations (Di Zio et al., 2007).

4.2.2 Categorisation of Missing Values

As we mentioned before in section 4.1.3 on page 67, Rubin categorized the missing value problem into three categories: missing completely at random (MCAR), missing at random (MAR), and not missing at random/missing not at random (NMAR/MNAR) (Little and Rubin, 2019; Di Zio et al., 2007; Thijs et al., 2002). The MAR assumption is generally used in clinical epidemiological research (Pedersen et al., 2017; Van der Heijden et al., 2006). It is critical to determine the category of the data, in order to choose a statistical strategy (Fielding et al., 2008; Moons et al., 2006).

4.2.3 Methods Used for Imputing Missing Values

To treat and estimate the missing values, Breiman (2001) proposed a non-parametric random forest (RF_m) model, which is an extended version of classification and regression trees (CARTs) and involves a supervised learning group method. The method used to build the trees involves replacement sampling of the main data set. The classification and regression trees are created using the training data bootstrap samples and tree induction using random feature selection (Svetnik et al., 2003; Bagheri et al., 2019). The performance of a tree is evaluated on the remaining data, which are contained in an out-of-bag sample.

Missing Imputation Using Random Forest (RF_m)

The best RF_m is determined based on the out-of-bag error, which is an unprejudiced gauge of the true prediction error (Shahn et al., 2015). RF_m has following advantages: (1) it is applicable even when number of variables is greater than the number of samples; (2) it is not prone to multicollinearity; (3) it is suitable for non-linear trends; (4) it does not suffer from the overfitting problem with an increase in the number of trees; and (5) it can tolerate outliers and missing values (Fan et al., 2019).

Missing Imputation Using Sequential Random Forest (missForest)

Another algorithm based on RF_m , called sequential random forest (missForest), has recently been developed for missing data imputation (Stekhoven and Bühlmann, 2012). This algorithm can impute missing values on any kind of data and its goal is the prediction of every single missing value, instead of drawing random values from a distribution. This algorithm can handle multivariate data sets concurrently comprising categorical and continuous variables (Shah et al., 2014). The key advantages of missForest over other imputation methods include: (1) having no requirement for the tuning of parameters; (2) it does not depend on assumptions pertaining to the distribution of data sets; (3) it allows for assessment of imputation quality without setting test data or laborious cross-validations using out-of-bag imputation error estimates; and (4) it provides above-par imputation results, even for high-dimensional data sets (i.e., when the number of variables is greater than the number of observations) (Stekhoven and Bühlmann, 2012).

In the present investigation, we consider four data mining techniques to predict the missing values in the data from the Kuwait Registry for Rheumatic Diseases (KRRD): predictive mean matching (PMM), k-nearest neighbours (kNN), random forest (RF_m), and sequential RF_m (missForest). The main objective of this study was to handle missing data in the KRRD, where the amount of missing data varied between 1% and 65.5% per variable (Table 4.5). Our secondary objectives were to choose the best missing data mechanism (MAR, MCAR, or MNAR) when assuming three different rates of missingness (10%, 20%, and 30%), as well as to compare the selected imputation methods (PMM, RF_m , kNN, and missForest) for each missing data mechanism under each rate of missingness. To select the best method for imputing missing data in KRRD, the root mean square error (RMSE) was used to evaluate the best imputation method which minimized the difference between the imputed data points and the original data points (that were subsequently set to missing).

Table 4.5: Study variables with abbreviations and with the percentages of missing values for each variable.

Variable name	Abbreviation	Measures	Missing rate	Variable role
RA Disease Duration		baseline	12.4%	independent
Smoking		baseline	26.0%	independent
Rheumatoid Factor	RF	baseline	8.3%	independent
Antinuclear Antibodies	ANA	baseline	21.4%	independent
Anti-Cyclic Citrullinated Peptide	ACPA	baseline	21.0%	independent
Sicca Symptoms	SICCA	baseline	19.8%	independent
Rheumatoid Nodules	Nodules	baseline	18.5%	independent
Family History	FH	baseline	28.4%	independent
Treatment Class	TC	repeated	13.7%	independent
Steroid Therapy	Steroid	baseline	6.6%	independent
Joint Pain		repeated	3.8%	independent
Disease Activity Score 28	DAS28	repeated	1.0%	target (outcome)
Erythrocyte Sedimentation Rate	ESR	repeated	5.1%	independent
C-Reactive Protein	CRP	repeated	2.2%	independent
Health Assessment Questionnaire Disability Index	HAQ	repeated	65.5%	independent

4.2.4 Data Source—Kuwait Registry for Rheumatic Diseases (KRRD)

All rheumatoid arthritis (RA) patients in this study were officially registered in the Kuwait Registry for Rheumatic Diseases (KRRD). The KRRD is a national registry listing adult patients with rheumatic diseases. Patients who fulfilled the American College of Rheumatology (ACR) criteria for RA (Aletaha et al., 2010) registered from January 2012 through March 2020 were included in the study. The RA information data were collected from the rheumatology departments of four major government hospitals in Kuwait, based on patient visits. The selected hospitals are distributed in different governorates covering the ethnic diversity of the Kuwaiti population. The KRRD, from which this study originated, was approved by the Ethics Committees of the Faculty of Medicine at Kuwait University and the Ministry of Health. Additionally, informed consent was obtained from all represented patients enrolled in the registry (Al-Herz et al., 2016).

Using the data obtained from KRRD, we conducted an in-depth comparative analysis of the different imputation methods. Missing data were entered into each data set,

assuming a general missing data pattern and three mechanisms of missing data: MCAR, MAR, and MNAR. Under the MCAR assumption, missing values were randomly applied to each data set. Under the MAR assumption, the probability of information being missing depended on class attribute. Under the MNAR assumption, the largest or smallest values of X_s were removed. The objective of the study was to derive a comparison of four different imputation methods for MNAR, MAR, and MCAR, concerning missing data. We simulated the rates of missing data by varying the proportions of missing values by 10%, 20% and 30%.

4.2.5 Calculating RA Indices

RA disease activity scores are measured using two different indices: DAS28 and CDAI. The DAS28 is the sum of four outcome parameters: TJC28, the number of tender joints (0–28); SJC28, the number of swollen joints (0–28); ESR, the erythrocyte sedimentation rate (in mm/h) (C-reactive protein (CRP) may be used as an alternative to ESR in the calculation); and GH, the patient global health assessment (from 0 = best to 100 = worst) (Equation 4.14).

$$DAS28 = 0.56 \times \sqrt{TJC28} + 0.28 \times \sqrt{SJC28} + 0.70 \times \ln(ESR \text{ Or } CRP) + 0.014 \times GH. \quad (4.14)$$

4.2.6 Multiple Imputation (MI) Process Using Rubin’s Rules

For our data sets, we used Rubin’s rules (Little and Rubin, 2019) for handling missing data. The MI process was conducted separately for each variable in the data set (figure 4.7). The first step in multiple imputation is to create values (*imputes* or m_i), with 5 iterations for each m_i (Imputed: set 1 to set 5, see figure 4.1) to be substituted for the missing data. To create the imputed values, we need to identify a model (e.g., a linear regression) that allows us to create imputes based on other variables in the data set (predictor variables). As we needed to perform this multiple times to produce multiple-imputed data sets, we identified a set of regression lines that were similar to

each other.

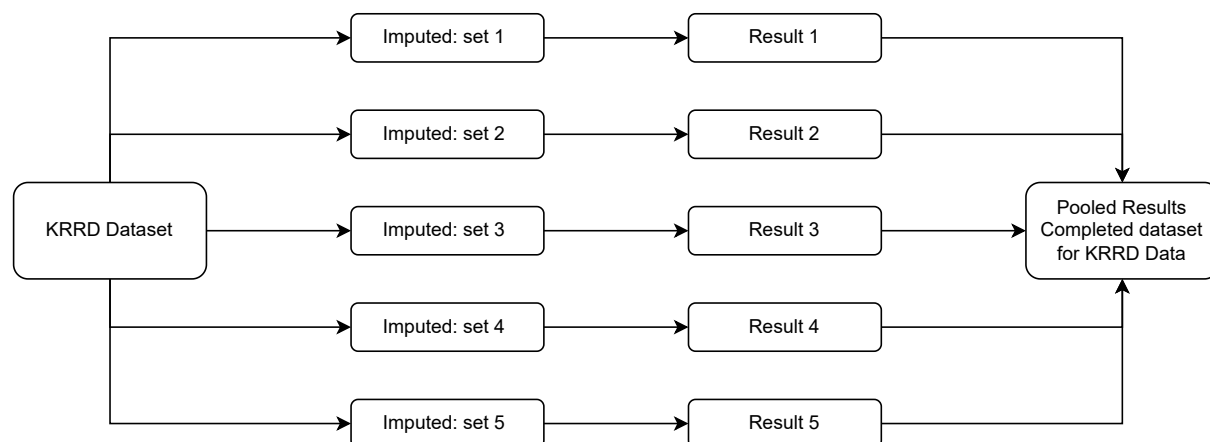


Figure 4.7: The steps of implementing multiple imputations using Rubin's rules to estimate missing values for the Kuwait Registry for Rheumatic Diseases (KRRD)

4.2.7 Number of Needed Imputations

An important aspect of previous technical treatments of multiple imputation is that the discussion of selecting the number of imputations that are required for acceptable statistical inference (e.g., (Rubin, 1987; Schafer, 1997; Schafer and Olsen, 1998)). For example, Schafer and Olsen (1998) recommend that in several applications, simply 3-5 imputations are enough to get sufficient results. Many are surprised by the claim that only 3-5 imputations may be needed. Rubin (1987) shows that the efficiency of an estimate based on m imputations is approximately

$$\left(1 + \frac{\gamma}{m}\right)^{-1} \quad (4.15)$$

where γ is the fraction of missing information for the quantity being estimated, so that the gains rapidly diminish after the first few imputations. In most situations there's merely very little advantage to generate and analyse over more than a few imputed datasets ("m"). In theory, the more imputation, the better performance in estimating missing values, but it takes a lot of time, which is a barrier for this research. It is convenient to set $m = 5$ during the stage of model building, and raise the amount in the

evaluation stage if it is needed (Van Buuren, 2018). So, in this study, the MI methods are performed with $m = 5$ imputed data sets which can be considered as satisfactory (Rubin, 1987).

4.2.8 Multiple Imputation Using RF_m Method

We described this approach before in 4.1.6 on page 73, and we propose using the random forest technique to impute missing observations. As it has previously been mentioned in section 4.1.6, the random forest algorithm has a built-in routine to handle the values that are missing. This is achieved by weighing the frequencies of values with the proximity of a random forest after the training of the mean data set is initially imputed (Cutler et al., 2012). This approach needs a response variable that is complete and useful for forest training. After imputing the missing values, the performance of different methods was assessed using the normalised root mean squared error (NRMSE) (Oba et al., 2003) for the continuous variables, defined by:

$$\text{NRMSE} = \sqrt{\frac{\text{mean} \left((\mathbf{X}^{\text{true}} - \mathbf{X}^{\text{imp}})^2 \right)}{\text{var} (\mathbf{X}^{\text{true}})}}, \quad (4.16)$$

where \mathbf{X}^{true} and \mathbf{X}^{imp} are the complete data matrix and the imputed data matrix, respectively. In this study, all predictors were classified as continuous observations. The mean and variance are used as a short notation for the empirical mean and variance computed over the missing values only, respectively. When an RF_m fits to the part that is observed on a variable, we reach the out-of-bag (OOB) estimate of the error for the variable. When the stopping criterion (γ) is met, we average it over the variable set of that type to obtain an approximation of the actual errors of imputation. We assessed the estimation performance by comparing the absolute difference between the OOB imputation error estimate in all simulation runs and the true imputation error.

4.2.9 Evaluation Criteria

The KRRD data set was simulated with these imputation methods; the best method was selected according to the RMSE score. To determine the best imputation method, three model performance measures were considered (Bennett et al., 2013): root mean square error (RMSE), mean absolute error (MAE), and correlation coefficient (R), which are respectively calculated as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (4.17)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (4.18)$$

where y_i and \hat{y}_i are the i^{th} observations for the comparison and reconstructed data sets, respectively. The error is measured based on the difference between the estimated and observed values. For RMSE and MAE, the smaller the value obtained, the more accurate the estimation method.

4.2.10 Study 2 Results

The performance of three imputation mechanisms (MCAR, MAR, and MNAR) was analysed using sub-data sets from KRRD patients using three different missingness rates (10%, 20%, and 30%). A total of 1,735 patients (62.8% women and 37.2% men) from 2012–2020 were included in this study (Table 4.6). The baseline investigated patient characteristics included factors such as smoking, RF, SICCA, ANA, ACPA, family history, treatment class, comorbidity, steroid and joint pain.

The average duration of RA disease was 9.19 ± 6.76 (SD) years. Most of the data were recorded at Amiri and Farwaniya hospitals (79%). A majority of the patients were non-smokers (89.6%), 77.1% were RF-positive, 65.9% of RA patients were ACPA-positive, and 58.8% had joint pain. The results showed a minority of RA patients had positive SICCA (18.7%), positive ANA (28.4%), positive family history (18.4%), and

positive steroid use (22.6%).

Table 4.6: Baseline patient characteristics of KRRD (2012 to 2020). In brackets is the percentage of cases or the standard deviation of the variable according to the type of the variable.

Variables	<i>n</i> or mean (% or <i>SD</i>)	N=1735
Sex (Female)	1,090 (62.8%)	1,735
Age (years)	54.0 (12.6)	1,719
RA Disease Duration (years)	9.19 (6.76)	1,520
Nationality		1,735
Kuwaiti	839 (48.4%)	
Non-Kuwaiti	896 (51.6%)	
Main Hospital		1735
Amiri	708 (40.8%)	
Farwaniya	663 (38.2%)	
Jahra	83 (4.78%)	
Mubarak	280 (16.1%)	
Sabah	1 (0.06%)	
Smoking (Yes)	133 (10.4%)	1,284
RF (Positive)	1,227 (77.1%)	1,591
SICCA (Yes)	260 (18.7%)	1,391
ANA (Positive)	388 (28.4%)	1,364
ACPA (Positive)	903 (65.9%)	1,370
Family History (Positive)	229 (18.4%)	1,243
Treatment Class (Biologics)	488 (32.6%)	1,498
Co-morbidity (Yes)	926 (53.4%)	1,735
Current Steroid (Yes)	366 (22.6%)	1,620
Joint Pain (Yes)	982 (58.8%)	1,670

Note: All categorical variables were described using frequencies and percentages (e.g. sex, nationality, main hospital, smoking, RF, SICCA, ANA, ACPA, family history, treatment class, co-morbidity, current steroid and joint pain). All the variables that are scale were described by mean and standard deviation (e.g. age and RA disease duration).

Table 4.7 provides a descriptive analysis of RA lab tests for ESR, CRP, HAQ, and DAS28, which were calculated five different times from five different data sets after implementing missing values' methods (original data set compared with imputed data sets using PMM, RF_m , kNN, and missForest).

The mean and SD value of ESR, CRP, HAQ, and DAS28 were 27.5729 ± 22.2706 , 5.9904 ± 4.9334 , 0.9517 ± 0.6649 , and 2.6756 ± 1.2902 , respectively, for the original data, and the mean and SD for the imputed data sets values ranged between 27.0639

Table 4.7: The mean and standard deviation for ESR, CRP, HAQ, and DAS28 from the original data set and the imputed data sets (IM).

Data Set	Variable	N	Minimum	Maximum	Mean	SE	SD
Original data	ESR	10703	0.0000	134.0000	27.5729	0.2153	22.2706
	CRP	8769	0.0000	21.0000	5.9904	0.0527	4.9334
	HAQ	4004	0.0125	3.0000	0.9517	0.0105	0.6649
	DAS28	11213	0.0000	9.7050	2.6756	0.0122	1.2902
$IM_1 = PMM$	ESR	11282	0.0000	134.0000	27.0701	0.2066	21.9426
	CRP	11282	0.0000	21.0000	6.4456	0.0441	4.6873
	HAQ	11282	0.0125	3.0000	0.9053	0.0044	0.4647
	DAS28	11282	0.0000	9.7050	2.6761	0.0121	1.2883
$IM_2 = RF$	ESR	11282	0.0000	134.0000	27.0639	0.2068	21.9637
	CRP	11282	0.0000	21.0000	6.4426	0.0442	4.6961
	HAQ	11282	0.0125	3.0000	0.9042	0.0044	0.4657
	DAS28	11282	0.0000	9.7050	2.6763	0.0121	1.2878
$IM_3 = kNN$	ESR	11282	0.0000	134.0000	27.1245	0.2074	22.0287
	CRP	11282	0.0000	21.0000	6.4323	0.0441	4.6865
	HAQ	11282	0.0125	3.0000	0.9061	0.0044	0.4658
	DAS28	11282	0.0000	9.7050	2.6767	0.0121	1.2876
$IM_4 = missForest$	ESR	11282	0.0000	134.0000	27.0939	0.2064	21.9236
	CRP	11282	0.0000	21.0000	6.4396	0.0440	4.6772
	HAQ	11282	0.0125	3.0000	0.9062	0.0044	0.4628
	DAS28	11282	0.0000	9.7050	2.6759	0.0121	1.2861

and 27.1245, 6.4323 and 6.4456, 0.9042 and 0.9062, and 2.6759 and 2.6767, respectively. The skewness and kurtosis values were mostly positive: 1.2248, 1.0182, 0.8120, and 0.5963 for skewness and 1.6256, 0.3932, -0.0311 , and 0.2599 for kurtosis in the case of imputed data sets that ranged between 1.2551 and 1.2675, 0.8085 and 0.8135, 1.1352 and 1.1430, and 0.5961 and 0.6005 for skewness, respectively, and 1.7394 and 1.7854, 0.1395 and 0.1493, 2.1178 and 2.1560, and 0.2664 and 0.2798 for kurtosis, respectively. The data showed that the original data set and the imputed data sets had very close values with small differences for all RA lab tests (ESR, CRP, HAQ, and DAS28).

4.2.11 Predicting the Influence of RA Factors on DAS28 Using the Original Data Set

Here we are trying to estimate a regression model that explains the effect from the independent variables (see Table 4.5) toward the outcome variable (DAS28). We used the original Kuwait Registry for Rheumatic Diseases (KRRD) dataset. As we mentioned before, the original dataset contains missing values in all variables (see Table 4.5).

The rate of missing values in the original data set vary from 2% to 66%. Table 4.8 shows the estimated parameters for predicting DAS28 using a multiple linear model. Only six variables were found to be significant risk factors that influence DAS28 ($R^2_{DAS28} = 0.773$). The results showed that ESR, CRP, HAQ, disease duration, and current steroid use were risk factors predicting DAS28, with $\beta = 0.034, 0.020, 0.129, 1.489, 0.247, 1.095$, respectively (95% CIs: 0.032–0.036, 0.012–0.029, 0.075–0.183, 1.415–1.562, 0.138–0.356, and 0.963–1.228, respectively). Other factors (RF, ANA, ACPA, SICCA, nodules, smoking, family history and joint pain) were not found to be risk factors influencing DAS28 (Table 4.8).

Because of the existence of missing values, smoking, joint pain and SICCA were not found to be a significant risk factor for DAS28. However, many scholars showed that those variables (smoking, joint pain and SICCA) can be risk factors toward DAS28 (e.g. Martínez et al. (2020) and Choe et al. (2013)).

4.2.12 Predicting the Influence of RA Factors on DAS28 from the PMM-Imputed Data Sets

Using the imputation process to predict all missing values in the KRRD data set using the three different missing imputation mechanisms (MAR, MCAR, and MNAR), we constructed a quality data set after fixing all missing values using PMM. Table 4.9 shows the estimated parameters when predicting DAS28 using multiple linear models and the PMM-imputed data sets.

The regression results showed the same significant risk factors, this time adding

Table 4.8: Multiple regression coefficients with 95% confidence intervals (in parentheses) for predicting DAS28 using the original data set including the missing values.

DAS28	
Data Set = Original	
ESR	0.034*** (0.032, 0.036)
CRP	0.020*** (0.012, 0.029)
HAQ	0.129*** (0.075, 0.183)
RF	0.021 (-0.064, 0.106)
ANA	-0.062 (-0.139, 0.014)
ACPA	0.008 (-0.066, 0.082)
SICCA	0.083 (-0.011, 0.178)
Nodules	-0.519 (-1.122, 0.083)
Smoking	0.184 (-0.019, 0.350)
Family History	-0.087 (-0.174, 0.001)
Joint Pain	0.273 (-0.015, 0.530)
Disease Duration	1.489*** (1.415, 1.562)
Current Steroid	0.247*** (0.138, 0.356)
Constant	1.095*** (0.963, 1.228)
R^2	0.773
Adjusted R^2	0.769

Note:

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

rheumatoid factor (RF) to predict DAS28 ($R^2_{DAS28} = 0.727$); in addition, smoking was found to be a very strong risk factor when we used the imputed data sets, but was not when we used the original data set to predict the influence on DAS28.

The regression model that used the original data set did not indicate that RF has a significant influence on DAS28 but, if we used the PMM-imputed data set, the regression model indicated that RF had a significant influence when predicting DAS28.

Due to the effect of bias induced by the missing values, the RF results were not significant when using the original data set; however, after we imputed all the missing data in the KRRD data set, the regression results became more sufficient and reliable.

Table 4.9: Multiple regression coefficients with 95% confidence intervals (in parentheses) to predict DAS28 from other predictors from PMM- and RF_m -imputed data sets.

<i>DAS28</i>		
	Imputed data set	
	(PMM)	(RF_m)
ESR	0.031*** (0.030, 0.031)	0.031*** (0.030, 0.031)
CRP	0.015*** (0.012, 0.017)	0.015*** (0.012, 0.017)
HAQ	0.202*** (0.178, 0.226)	0.202*** (0.178, 0.225)
RF	0.061*** (0.035, 0.087)	0.050*** (0.024, 0.076)
ANA	0.005 (-0.020, 0.031)	0.003 (-0.023, 0.028)
ACPA	0.003 (-0.020, 0.026)	0.008 (-0.016, 0.031)
SICCA	0.064*** (0.034, 0.094)	0.060*** (0.031, 0.090)
Nodules	0.016 (-0.044, 0.077)	-0.011 (-0.071, 0.049)
Smoking	0.131*** (0.086, 0.177)	0.140*** (0.095, 0.186)
Family History	-0.029 (-0.059, 0.001)	-0.022 (-0.052, 0.008)
Joint Pain	0.674*** (0.662, 0.686)	0.676*** (0.664, 0.688)
Disease Duration	-0.007*** (-0.009, -0.006)	-0.007*** (-0.009, -0.005)
Current Steroid	0.128*** (0.093, 0.162)	0.118*** (0.084, 0.153)
Constant	-0.032 (-0.146, 0.083)	0.002 (-0.112, 0.117)
Observations	11,282	11,282
R^2	0.727	0.728
Adjusted R^2	0.727	0.728

Note:

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

4.2.13 Predicting the Influence of RA Factors on DAS28 from the Imputed Data Sets Using kNN

The kNN-based imputation of missing data restored the values of observations for neighbouring data points in the data set (e.g. the KRRD data set) and better standard data were obtained (Malarvizhi and Thanamani, 2012). As shown in Table 4.10, various parameters were used to establish DAS28 prediction using the kNN-imputed data sets.

The regression results were calculated based on similar factors as those of the original data set with further incorporation of RF and SICCA and further prediction of DAS28 ($R^2_{DAS28} = 0.727$).

Similar to the PMM-based analysis, the kNN imputation revealed that RF can significantly influence the prediction of DAS28, which was not significant in the original data set due to bias. The disease duration factor had a negative value in both PMM

and kNN, whereas it was positive in the original data set.

4.2.14 Predicting the Influence of RA Factors on DAS28 from the RF_m -Imputed Data Sets

The results of RF_m imputation, in terms of removing bias in the KRRD data set, were similar to those of PMM-based imputation. Table 4.9 shows that the factors that significantly affected the DAS28 parameter at $p < 0.01$ were similar between PMM- and RF_m -based imputation, with their values being very close. The adjusted R_{DAS28}^2 value of 0.728 was obtained after imputation.

4.2.15 Predicting the Influence of RA Factors on DAS28 from the missForest-Imputed Data Sets

One of the best methods for imputation reported in the literature and evident from the analysis was missForest. A part from all the factors listed in the original data set and compared to the imputation by kNN, and RF_m , the missForest-based imputation analysis produced better results, as evidenced by MAR, MCAR, and MNAR missing value mechanisms (Table 4.10).

The adjusted R_{DAS28}^2 value was 0.731. We conclude, based on the better results, that the data set was the most refined and its quality was the most improved after applying the missForest imputation method.

We hypothesized that the missing data could be imputed using the different imputation strategies; therefore, the MCAR, MAR, and MNAR mechanisms were simulated for missing values in the three different missingness proportions of 10%, 20%, and 30%. As shown in Table 4.11, the RMSE value ranged between 2.518 to 6.066 for MAR, 2.555 to 5.590 for MCAR, and 3.631 to 8.004 for MNAR. The MAR had the lowest RMSE, compared to the other missing data methods.

Similar investigations have been previously performed and our results were in agreement with those in the earlier reports (Valdiviezo and Van Aelst, 2015; Junger and De Leon, 2015).

Table 4.10: Multiple regression coefficients with 95% confidence intervals (in parentheses) to predict DAS28 from other predictors from kNN- and missForest-imputed data sets.

<i>DAS28</i>		
Imputed data set		
	(kNN)	(missForest)
ESR	0.031*** (0.030, 0.031)	0.031*** (0.030, 0.031)
CRP	0.015*** (0.013, 0.018)	0.015*** (0.012, 0.017)
HAQ	0.203*** (0.180, 0.227)	0.204*** (0.180, 0.227)
RF	0.058*** (0.032, 0.084)	0.056*** (0.030, 0.082)
ANA	0.001 (−0.025, 0.026)	0.008 (−0.018, 0.033)
ACPA	0.004 (−0.019, 0.027)	0.004 (−0.020, 0.027)
SICCA	0.065*** (0.035, 0.094)	0.057*** (0.027, 0.087)
Nodules	−0.002 (−0.062, 0.058)	0.004 (−0.056, 0.063)
Smoking	0.132*** (0.087, 0.177)	0.139*** (0.093, 0.184)
Family History	−0.024 (−0.054, 0.006)	−0.021 (−0.051, 0.010)
Joint Pain	0.674*** (0.662, 0.686)	0.677*** (0.666, 0.689)
Disease Duration	−0.007*** (−0.009, −0.005)	−0.007*** (−0.009, −0.005)
Current Steroid	0.125*** (0.091, 0.159)	0.129*** (0.094, 0.163)
Constant	−0.017 (−0.131, 0.098)	−0.037 (−0.151, 0.077)
Observations	11,282	11,282
R^2	0.728	0.731
Adjusted R^2	0.728	0.731

Note:

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

In the MAR, MCAR, and MNAR mechanisms, missForest was the best method of imputation, having the lowest RMSE values for all of the parameters and at all three percentages of simulated missing data (MAR: 2.518, 3.013, and 3.032; MCAR: 3.168, 2.555, and 2.871; and MNAR: 4.962, 4.180, and 3.631 for 10%, 20%, and 30%, respectively) these results agreed with Kokla et al. (2019); Tang and Ishwaran (2017); Valdiviezo and Van Aelst (2015) and Stekhoven and Bühlmann (2012).

This was followed by kNN, which performed better than the other two imputation methods (RF_m and PMM), in terms of RMSE values, at every percentage of missingness (MAR: 4.107, 4.884, and 4.184; MCAR: 3.820, 3.560, and 3.734; and MNAR: 6.236, 5.507, and 5.062 for 10%, 20%, and 30%, respectively); see Table 4.11. Similar results have been reported that strongly support the better imputation of kNN, compared with RF_m and PMM (Zakaria and Noor, 2018). RF_m and PMM were the worst-performing

Table 4.11: Comparison between imputation methods after we simulated 10%, 20%, and 30% missing data in the KRRD data set. The RMSE is used to highlight and select the best missing imputation method with the lowest RMSE score.

Method	MAR			MCAR			MNAR		
Missingness rate	10%	20%	30%	10%	20%	30%	10%	20%	30%
Predictive mean matching (PMM)	5.349	6.066	4.944	5.590	4.590	5.135	6.950	7.471	6.516
Random forest (RF_m)	4.618	5.233	5.234	4.837	4.204	4.539	8.004	7.212	6.737
Classification and regression trees (kNN)	4.107	4.884	4.184	3.820	3.560	3.734	6.236	5.507	5.062
missForest	2.518	3.013	3.032	3.168	2.555	2.871	4.962	4.180	3.631

multiple imputation methods; of the two, using RF_m had a slight advantage over PMM but PMM had better imputation in a few of the cases, such as MAR 30% or MNAR 10% and 30%, where RF_m had a larger RMSE value than PMM. Table 4.9 and Table ?? represent the multiple regression coefficients with 95% confidence intervals (CIs) for the prediction of DAS28 using the imputed data sets (PMM, RF_m , kNN, and missForest). The table demonstrates the effect of patient demographics on RA disease activity, where DAS28 was the response variable.

The disease activity score for DAS28 is also reported; where $R^2_{DAS28} = 0.727$ for PMM method, and $R^2_{DAS28} = 0.728$ for RF_m method. Regarding kNN and missForest, $R^2_{DAS28} = 0.728$ for kNN method, and $R^2_{DAS28} = 0.731$ for missForest method. The results show the positive effect of various factors, such as ESR, CRP, HAQ, RF, SICCA, smoking, joint pain, and current steroid use, with $\beta = 0.031, 0.015, 0.202\text{--}0.204, 0.050\text{--}0.061, 0.057\text{--}0.065, 0.131\text{--}0.140, 0.674\text{--}0.677$, and $0.118\text{--}0.129$, respectively, on RA disease activity, whereas family history and disease duration—with $\beta = (0.029)$ to (-0.021) and 0.007 , respectively—had negative effects under all four imputation methods (Table 4.9 and Table 4.10).

Additionally, nodules showed diverse effects per imputation method. The nodules had positive values for PMM and missForest, with $\beta = 0.016$ and 0.004 , respectively, and negative values for RF_m and kNN, with $\beta = -0.011$ and -0.002 , respectively. The constant had negative values for PMM, kNN, and missForest, with $\beta = -0.032, -0.017$, and -0.037 , respectively, and a positive value for RF, with $\beta = 0.002$ (Table 4.9 and

Table 4.10).

4.2.16 Study 2 - Discussion and Conclusion

The obtained rheumatoid arthritis (RA) patient data recorded in the Kuwait Registry for Rheumatic Diseases (KRRD) registry were utilised to quantify the Rheumatoid Arthritis Disease Activity Score. All the information was acquired from 1,735 patients from public healthcare facilities with permission from the relevant ethical committees. The baseline variables under investigation for every patient included smoking, sex, disease duration, age, nationality, SICCA, RF, ACPA, ANA, family history, treatment class (biologics, cDMARDs), current steroids, comorbidity, DAS28 group, and joint pain.

Systematic errors that existed between the anticipated and noted values due to the missing value led to outcome bias. However, to get the accurate missing values (Alsaber et al., 2020), it is important to eliminate the bias and apply the optimal approach to guarantee reliability and quality of data analysis. The uncompleted data sets contradicted significantly with the complete data file (Forbes et al., 2004). The emergence of imputation algorithms has been attributed to their substantial global use.

Imputation methods overcome the existing bias caused by missing values. However, their values may potentially lead to bias in the result. Therefore, they should be used vigilantly. The research utilised numerous variables to define the RA disease activity scores. The application of many factors resulted in data with missing characteristics in various patients, leading to biased results. The focal point was to identify the suitable imputation approach to complete the missing features in the RA data set.

The variation of the percentage of missing data ranged from 2% to 66%. At this point, four imputation approaches were assessed, the kNN, PMM, missForest, and RF_m throughout three diverse missing approaches MAR, MNAR, and MCAR with Kuwait Registry for Rheumatic Diseases (KRRD) RA infection data set. Performance evaluations of the imputation methods were done utilising RMSE values, with the minimum RMSE value showing the best imputation technique.

Multiple imputations (MI) are computationally comprehensive and require estima-

tions. To get enough needed results, several algorithms should be run frequently, where running time increases with more missing data. Concerning our missing data problem, we conclude that MI imputation using missForest is the most efficient approach for our research. This has been a successful adapted approach in our data set that is related to the medical field as it provides an attractive balance of both accuracy and conceptual simplicity. However, the balance of the statistical expertise of the research team, validity of the method, and ease of interpretability for readers must be taken into account in order for the optimal imputation method to be deemed successful. MI outperformed the single imputation methods or deletion (Shrive et al., 2006).

The findings, in this case, are similar to the findings obtained when using a hypothetical data set to differentiate missing data approaches. The current research was possible since the variable estimates attained by each missing data approach could be contrasted to the already known values of the variables of an absolute data set acquired from the clinical setting. The result reveals that is possible to apply missing data methods like MI in the current context (Baraldi and Enders, 2010).

However, despite the effectiveness of the MI method in replacing the missing data, it is significant to note that the associated problem with missing data cannot be improved by any missing data approach. MI and numerous missing data techniques are useful for MAR or MCAR despite their unreliability when data is MNAR. Determination of whether data is MAR or MNAR is often difficult as there is no reliable technique to do so. But, in some clinical or environmental studies (e.g. Tsiampalis and Panagiotakos (2020); Alsaber et al. (2021b); Mishra and Khare (2014); and Pedersen et al. (2017)), either MAR or MCAR are preferable rather than the MNAR mechanism.

Finally, the variety of data and the negative effects of missing data, and the correlated variables that come with using traditional approaches to handle the missingness of data are not considered important (Baraldi and Enders, 2010). The findings show that techniques like MI perform better than the traditional approaches as they facilitate the reintroduction of the difference that would occur upon attaining missing scores (i.e. multiple imputation better handles missing data by estimating and replacing missing

values many times). Multiple imputation fills in missing values by generating plausible numbers derived from distributions of and relationships among observed variables in the data set (Rubin, 1987).

Multiple imputation differs from single imputation methods because missing data are filled in many times, with many different plausible values estimated for each missing value. Using multiple plausible values provides a quantification of the uncertainty in estimating what the missing values might be, avoiding creating false precision (as can happen with single imputation). Multiple imputation provides accurate estimates of quantities or associations of interest, such as treatment effects in randomised trials, sample means of specific variables, correlations between two variables, as well as the related variances. In doing so, it reduces the chance of false-positive or false-negative conclusions. As a result, this reduces bias produced by missing data and enhances the ability to realise meaningful results. MI and other techniques are fast and elementary to use and their long term merits are worth the time taken to learn the techniques and apply it within the clinical research setting. From the results, missForest is regarded as the most productive imputation technique with the least RMSE, reproduced employing 10%, 20%, and 30% missing data. The RF_m and PMM were identified as the worst performing imputation techniques. Due to the availability of large data from the registered RA patients used, the research and its outcomes are considered robust. Additionally, the imputation method considered and missingness procedures (implemented at 10% to 30%, utilising MAR, MNAR, and MCAR) improved data reliability with notable p-values attained (Li et al., 2015). MissForest is a highly accurate method of imputation for missing data in KRDD data sets and outperforms other common imputation techniques in terms of imputation error and maintenance of predictive ability with imputed values in clinical predictive models. This approach can be used in data registries to improve the accuracy of data, including the ones for rheumatoid arthritis patients.

Chapter 5

Time Series Statistical Methods

Review

5.1 Introduction

This chapter gives an overview of the time series analysis procedures and the concepts employed for examining the long- and short-run relationship between air pollution and chronic disease activity among patients from Kuwait. It will also present the primary methods of time series analysis implemented in this thesis (i.e., univariate and multivariate time series analysis). We begin with a discussion of the fundamental concepts of the time series approach, followed by an examination of correlation- and partial-correlation matrices used. Next, the methodology for implementing a time series analysis is introduced, beginning with testing stationarity for each variable using the following methods: a) the Augmented Dickey-Fuller test (Dickey and Fuller, 1979a) the Phillips-Perron test (Phillips and Perron, 1988), and c) the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Dickey and Fuller, 1979b). Then, we present the ARIMA model to measure the time series association between chronic disease activity and the pollutants. The auto-arima model is used to set up the best ARIMA (p, d, q) by using the Akaike information criterion (AIC), and Schwarz Bayesian information criterion (SBC). The model is checked

by indicators: mean absolute percentage error (MAPE) and root mean square error (RMSE). In addition, the GARCH model is provided for modelling both a more pliable lag structure and a longer memory to explain the relationship between chronic disease activity and the effect of air pollutants.

Then, we switch from univariate time series modelling to a multivariate time series approach. As usual, we start with the Vector Autoregression (VAR), and the vector error-correction model (VECM). However, to implement the VECM, we have to evaluate the Granger Causality in a VAR. The lag length selection using information criteria will be presented in section 5.9.2 using three different criterion tools to establish the number of lags that should be added as regressors. The method of Johansen and Juselius' cointegration test will be mentioned. The VECM used in this thesis will be presented in section 5.15 after testing for the number of cointegration ranks in section 5.14. Finally, the impulse response function is implemented to trace the ramification of one standard deviation disturbance to one of the dependent variables on present and future estimates of the endogenous variables. The variance decomposition, as described in section 5.19, decomposes variation in an endogenous variable into component disturbances to the endogenous variable that exists in VAR.

5.2 Fundamental Concepts

5.2.1 Time Series Modelling

In the last few years, researchers have put significant effort into building and improving appropriate time series prediction models. Time series modelling seeks to model the process that produces time series data so that many statistical aspects of the observed data can be reproduced (Granger et al., 1986). Time series analysis involves the procedure of fitting a time series to the correct model (Juselius, 2006). Once an appropriate model is tailored to a time series, the associated parameters can be calculated using known data.

Later, the model is used to predict future events. This can then be used in forecasts and simulations, and it comprises various ways for trying to comprehend the series'

nature. There are two types of time series processes: stationary and non-stationary.

5.2.2 Time Series Definition

A time series refers to a collection of sequential data points measured in chronological order over a period of time, or a method that analyse the components and explainable parts of a time series, allowing for the detection of trends, estimations, and projections. The mathematical definition is a set of vectors $x(t), t = 0, 1, 2, \dots$, where t is the elapsed time (Cochrane, 2005; Hipel and McLeod, 1994; Raicharoen et al., 2003). Another way to say this is that it is the process of fitting a time series to an appropriate model (Hipel and McLeod, 1994), and then calculating associated parameters using known data values.

Fundamentally, time series analysis uses a model to estimate future values founded on known previous values to try to understand the underlying context of the data points. To forecast with a time series, previous observations are gathered and examined to construct a mathematical model reflecting the series' fundamental data generation process (Zhang, 2007, 2003). It can take several decades to build and improve suitable time series predictive models, which are then used to predict future events.

5.2.3 Time Series and Stochastic Process

In nature, a time series is usually non-deterministic, which means we cannot forecast what will happen in the future with certainty. A time series $\{x(t), t = 0, 1, 2, \dots\}$, is usually considered to follow a probability model (Cochrane, 2005) that defines the joint distribution of the random variable x_t . The stochastic process is a mathematical phrase used to explain the probability structure of a time series (Hipel and McLeod, 1994). As a result, the series' sequence of observations is a sample realisation of the stochastic process that created it. The time series variables X_t are usually assumed to be independent and identically distributed (*i.i.d*) in a normal distribution. However, time series are not strictly *i.i.d*; rather, they follow an approximately regular pattern throughout time (Cochrane, 2005). For example, if a city's temperature is extraordinarily high today, it may be reasonably assumed that the temperature tomorrow will be similarly high. This is why, when done correctly, time series forecasting produces results that are near to the real value. The choice of an appropriate model is critical, as it reflects the series' fundamental arrangement, and the fitted model is then utilised for subsequent forecasting.

Several models for time series analysis exist, including: autoregressive (AR), moving averages (MA), autoregressive moving averages (ARMA), autoregressive integrated moving averages (ARIMA), autoregressive conditional homoscedasticity (ARCH), generalised ARCH (GARCH), component GARCH, exponential GARCH, fractionally integrated GARCH, and threshold ARCH (Wu et al., 2012; Yusof et al., 2013; Sparks and Yurova, 2006). The focal point of this study is on the first four. Time series models can also be linear or non-linear, depending on whether the present value of the series is a linear or non-linear function of prior observations. These primary concepts and definitions of time series analysis, as well as the conditions, assumptions, processes, and principles included in the use of AR, MA, ARMA, and ARIMA, are thoroughly examined in this chapter.

$\mathbf{Z}_t = [Z_{1,t}, Z_{2,t}, \dots, Z_{m,t}]'$ is a stationary m-dimensions vector time series process if all of its component series are a univariate stationary process and its first two moments

are time-invariant. A stationary vector time series model or process is defined by its correlation matrix function, partial correlation matrix function, and mean vector. In contrast, a univariate stationary model or process is defined by its moments, such as its autocorrelation function, partial autocorrelation function, and mean, which we turn to now.

5.2.4 Correlation and Partial Correlation Matrix Functions

Let $\mathbf{Z}_t = [Z_{1,t}, Z_{2,t}, \dots, Z_{m,t}]'$, $t = 0, \pm 1, \pm 2, \dots$ be a stationary real-valued vector process of m -dimensions, with $E(Z_{i,t}) = \mu_i$ being constant for all $i = 1, 2, \dots, m$ and the cross-covariance between $Z_{i,t}$ and $Z_{j,s}$ being constant for all $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, m$ are functions only of the time difference $(s - t)$, where the mean vector is:

$$E(\mathbf{Z}_t) = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix} \quad (5.1)$$

and the covariance matrix with lag k

$$\begin{aligned} \boldsymbol{\Gamma}(k) &= \text{Cov}\{\mathbf{Z}_t, \mathbf{Z}_{t+k}\} = E\left[(\mathbf{Z}_t - \boldsymbol{\mu})(\mathbf{Z}_{t+k} - \boldsymbol{\mu})'\right] \\ &= E \begin{bmatrix} Z_{1,t} - \mu_1 \\ Z_{2,t} - \mu_2 \\ \vdots \\ Z_{m,t} - \mu_m \end{bmatrix} [Z_{1,t+k} - \mu_1, Z_{2,t+k} - \mu_2, \dots, Z_{m,t+k} - \mu_m] \\ &= \begin{bmatrix} \gamma_{1,1}(k) & \gamma_{1,2}(k) & \cdots & \gamma_{1,m}(k) \\ \gamma_{2,1}(k) & \gamma_{2,2}(k) & \cdots & \gamma_{2,m}(k) \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{m,1}(k) & \gamma_{m,2}(k) & \cdots & \gamma_{m,m}(k) \end{bmatrix}, \end{aligned} \quad (5.2)$$

where

$$\gamma_{i,j}(k) = E (Z_{i,t} - \mu_i) (Z_{j,t+k} - \mu_j). \quad (5.3)$$

for $k = 0, \pm 1, \pm 2, \dots$, $i = 1, 2, \dots, m$, and $j = 1, 2, \dots, m$. $\Gamma(k)$ is the covariance matrix function for the vector process \mathbf{Z}_t as a function of k .

The cross-covariance function where $i = j$, $\gamma_{i,i}(k)$ is the autocovariance function for the i th component process, $Z_{i,t}$; and if $i \neq j$, $\gamma_{i,j}(k)$ (i.e. between the series $Z_{i,t}$ and $Z_{j,t}$). The process's contemporaneous variance-covariance matrix is easily identified by the matrix $\Gamma(0)$.

5.3 Stationary Time Series

Stationary time series processes involve procedures with statistical outcomes, such as variance and mean, that are time insensitive. It is essential that time series models are shaped within the realm of possibility so that researchers may utilise them to estimate future values. This also allows the creation of models with minimised mathematical intricacy. There are two types of stationary processes, strictly (or strongly) stationary and weakly stationary. A process $\{x(t), t = 0, 1, 2, \dots\}$ is strictly stationary if the joint probability distribution function of $\{x_{t-s}, x_{t-s+1}, \dots, x_t, \dots, x_{t+s-1}, x_{t+s}\}$ is independent of t for every s . As a result, any feasible combination of random variables from a strictly/strongly stationary process has a joint distribution that does not depend on time. However, strict stationarity is not necessarily required in practical applications, thus a slightly weaker variant is proposed. A process is weakly stationary if the statistical moments of a stochastic process up to that order depend only on time differences of events of the data being used to approximate the moments, and not on the time itself. It is worth noting that neither weak stationarity nor strong stationarity have any implications on the other. A weakly stationary process that follows a normal distribution, for example, is also strongly stationary. To find stationarity in a time series data set, some mathematical tests, such as Dickey and Fuller's (Dickey and Fuller, 1979a), are

commonly utilised.

The mean of a stationary series is well-defined, and it can fluctuate with constant limited variation around it. For all t and j , a process is stationary if its first and second moments are time invariant: $E(y_t) = \mu$, $\text{Var}(y_t) = \sigma_0^2$ and $\text{Cov}(y_t, y_{t-j}) = \sigma_j^2$ for all t and j , where t is the observation date and j is the time interval between observations (Hamilton, 1994). The covariance between y_t and y_{t-j} is only determined by j , the time interval between observations, and not by t , the observation date.

Autoregressive models have been found to accurately reflect a variety of time series processes. The present value of the process, y_t , is shown in AR models as a linear combination of a white noise shock term ε_t and prior values of the process that are finite in nature. A univariate AR model of order 1 (AR1), for example, is defined as:

$$y_t = \phi_1 y_{t-1} + \varepsilon_t, \quad t = 1, 2, \dots, T. \quad (5.4)$$

Equation (5.4) can be written as $(1 - \phi_1 L) y_t = \varepsilon_t$ using the lag operator L . The root of the characteristic polynomial has to be on the exterior part of a unit circle for Equation (5.4) to be a stationary process, as established by Box and Jenkins (1976, pp. 47-82). That is, the outcome for $1 - \phi_1 z = 0$ must meet the condition $|z| > 1$. This means that the AR(1) in Equation (5.4) is stationary strictly if $|\phi_1| < 1$.

These stationarity factors can be extended to include multivariate time series processes. Consider time series vector \mathbf{Z}_t with k dimensions. VAR(p), the p^{th} -order vector autoregressive model for \mathbf{z}_t , is expressed as follows:

$$\mathbf{z}_t = \mathbf{\Pi}_1 \mathbf{z}_{t-1} + \mathbf{\Pi}_2 \mathbf{z}_{t-2} + \dots + \mathbf{\Pi}_p \mathbf{z}_{t-p} + \varepsilon_t, \quad t = 1, 2, \dots, T. \quad (5.5)$$

In this case $\mathbf{\Pi}_i$ is a $k \times k$ autoregressive coefficient matrix, $i = 1, 2, \dots, p$, and ε_t is a $k \times 1$ unobservable zero mean white noise vector process with covariance matrix $\mathbf{\Sigma}$. When the roots of $\det(\mathbf{I}_k \lambda^p - \sum_{i=1}^p \mathbf{\Pi}_i \lambda^{p-i}) = 0$ (the eigenvalue) lie inside the unit circle for all values of p , the VAR(p) in Eqn. (5.5) is stationary (Hamilton, 1994).

A MA-eigenvalue matrix's diagonal is the reciprocal of the corresponding root of the

characteristic polynomial, z , as established by Johansen (1996, pp. 15-16). As a result, all values of z meeting Eqn. (5.5) satisfy the alternative stationarity requirement that all values of z satisfying $\det(\mathbf{I}_k - \sum_{i=1}^p \mathbf{\Pi}_i z^i) = \det(\mathbf{\Pi}^*(z)) = 0$ should lie outside a unit circle (for $\lambda = 1/z$).

5.4 Non-Stationary Time Series

Most climate variables, including global greenhouse gas levels, mean temperatures, anthropogenic aerosols and solar irradiance, have shown increasing trends over the previous 150 years, making them non-stationary (Stern and Kaufmann, 2000; Stock and Watson, 2001; Liu and Rodri guez, 2005; Stern and Kaufmann, 1999; Kaufmann et al., 2006). It is also well-known that statistical procedures like regression analysis can produce erroneous conclusions when dealing with non-stationary data (Granger and Newbold, 1974). This could also happen in climate change attribution and detection studies that use static regression approaches to attribute and detect changes.

Rather than utilising time series models for climate to explain optimum variation in observed time series, it may be more useful to see if the time series climate model can remove a trend from the series under observation. Suitable effective time series models, such as vector autoregression (VAR) models, have the capability to handle non-stationarity appropriately and prevent erroneous attribution. Causality between climate variables can also be tested using such models (Granger, 1969). The statistics of non-stationary time series are dependent on the time period chosen. There are major tendencies in these systems that do not revert to the mean. If $\phi_1 = 1$ in Equation (5.5), the second moment of the process becomes a rising function of time, i.e., $\text{Var}(y_t - y_0) = t\sigma^2$, where σ^2 represents the variance of the error term ε_t . Removing trends or differencing can make a non-stationary time series stationary. Non-stationary processes are classified as either difference-stationary or trend-stationary. It is possible to model a series as a deterministic trend model, for example:

$$y_t = \mu + \theta t + \varepsilon_t. \quad (5.6)$$

Stationarity is achieved by deleting the linear deterministic trend $\mu + \theta t$; the zero mean white noise process is represented by ε_t . This strategy, on the other hand, would not be effective for a series produced by a stochastic trend model,

$$\begin{aligned} y_t &= \mu + y_{t-1} + \varepsilon_t, \\ &= \mu t + y_0 + \sum_{s=0}^t \varepsilon_s. \end{aligned} \quad (5.7)$$

When the deterministic factor is removed, the model is left with a cumulated total of error terms that is still not stationary. To produce a stationary process $\Delta y_t = \mu + \varepsilon_t$ for such a process, first differencing (5.7) is needed. An integrated process of order d is a non-stationary process that can only be made stationary after differencing d times and is indicated by $I(d)$. When the level stationarity is zero, it is a specific instance of trend-stationarity (5.7). If the VAR(p) process in (5.6) has a unit root, that is, if $z = 1, \det(\mathbf{I}_k - \sum_{i=1}^p \mathbf{\Pi}_i) = \det(\mathbf{\Pi}^*(1)) = 0$. As a result, the matrix $\mathbf{\Pi}^*(1)$ is non-invertible for a non-stationary VAR process.

As a result, the presence and type of trend in a time series should be determined before proceeding with further analysis. This is an important step since mistreating the type of trend can lead to seriously misleading results (for example, false removal of trends, false differencing, false regression, and so on).

5.5 Testing for Difference-Stationarity

Many different types of unit root test methodologies are useful in model construction (e.g., Phillips and Perron (1988); Dickey and Fuller (1981); Perron (1990); Dickey and Fuller (1979b); Choi et al. (2002)), however, the majority of them receive criticism for their lack of power in finite sample investigations (Cochrane, 1991; Faust, 1996; Blough, 1992). A reasonable strategy to minimize such hazards is an amalgamation of tests with

the addition of contrasting null hypotheses. When testing for non-stationarity of a time series, this thesis employs the augmented Dickey-Fuller non-stationarity test (Dickey and Fuller, 1981) and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) stationarity test (Dickey and Fuller, 1979b), which are discussed in fuller detail. These two methods were chosen since they both test contrasting null hypotheses, both have significant research supporting them, and are both widely used in the literature.

The stationarity of a series having the ability to impact its behaviour is a significant phenomenon. Modelling the x and y association as a basic OLS relationship, as in equation (5.8), will result in a false regression if the x and y series are non-stationary random processes (combined).

$$Y_t = \alpha + \beta X_t + \varepsilon_t. \quad (5.8)$$

Time series stationarity refers to the statistical features of a series over time, such as its variance and mean. The time series is considered a stationary process (that is, it is not a random walk or it does not have unit root) if they are all constant across time; otherwise, the series will be defined as a random walk or a series that has unit root (see equation (5.9)).

x level	x_t	
x 1 st -differenced value	$\nabla x_t = x_t - x_{t-1}$	(5.9)
x 2 nd -differenced value	$\nabla^2 x_t = (\nabla x_t - \nabla x_{t-1}) = x_t - 2x_{t-1} + x_{t-2}$	

A series that is stationary without differencing is referred to by $I(0)$, or integrated of order 0. A series that is stationary with first differences, on the other hand, is referred to as $I(1)$, or integrated of order one (1). The Phillips-Perron test and the enhanced Dickey-Fuller test (1979) will be implemented to determine whether the variables were stationary (Phillips and Perron, 1988; Dickey and Fuller, 1979a).

5.6 Unit-Root Test

When utilising an estimated model with non-stationary variables, false results can occur, and conclusions cannot be trusted. This is known as a false regression. Performing a unit-root test on every variable before the analysis can minimize false regression.

To find the unit root, we employed the Augmented Dickey-Fuller (ADF) test (Said and Dickey, 1984), and the Phillips and Perron (1988) test (PP test). One advantage of the PP test compared to the ADF test is that it does not need the assumption of homoscedasticity in the error term. The DF-GLS test outdoes the ADF test in terms of power and small sample size and power (Elliott et al., 1992).

5.6.1 Dickey-Fuller Tests (DF Test and ADF Test)

One of the most well-known and extensively utilised unit root tests is the Dickey-Fuller test (Dickey and Fuller, 1979a). It is founded on the first-order autoregressive process model (Box et al., 2015):

$$y_t = \phi_1 y_{t-1} + \varepsilon_t, \quad t = 1, \dots, T. \quad (5.10)$$

ε_t is the non-systematic constituent of the model that meets the features of the white noise process, where ϕ refers to the auto-regression parameter. The null hypothesis is $H_0 : \phi_1 = 1$, which states that the process has a unit root and is thus non-stationary, and is expressed as $I(1)$; the alternative hypothesis is $H_1 : |\phi_1| < 1$, which states that the process is stationary since it does not contain a unit root, and is expressed as $I(0)$. We utilise an expression that we get if we subtract y_{t-1} from both sides of the equal sign of the equation (5.10) to compute the test statistic for DF tests:

$$\Delta y_t = \beta y_{t-1} + \varepsilon_t. \quad (5.11)$$

where $\beta = \phi_1 - 1$. The test statistic is expressed as:

$$t_{DF} = \frac{\hat{\phi}_1 - 1}{s_{\hat{\phi}_1}}. \quad (5.12)$$

where $\hat{\phi}_1$ is a least squares estimate of ϕ_1 and $s_{\hat{\phi}_1}$ its standard error estimate. This test statistic follows the Dickey-Fuller distribution if the null hypothesis $\phi_1 = 1$ is true. Critical values for this distribution were derived by simulation and tabulated in Dickey (1976) and Fuller (2009).

A linear trend or a constant can be used to expand Model (5.10):

$$\begin{aligned} y_t &= \beta_0 + \phi_1 y_{t-1} + \varepsilon_t, \\ y_t &= \beta_0 + \beta_1 t + \phi_1 y_{t-1} + \varepsilon_t. \end{aligned} \quad (5.13)$$

The Augmented Dickey-Fuller test is developed when a non-systematic component in DF models is autocorrelated (Dickey and Fuller, 1981). After that, model (5.10) is changed into:

$$y_t = \phi_1 y_{t-1} + \sum_{i=1}^{p-1} \gamma_i \Delta y_{t-i} + \varepsilon_t. \quad (5.14)$$

The test statistic for the ADF test is calculated using the following equation:

$$\Delta y_t = (\phi_1 - 1) y_{t-1} + \sum_{i=1}^{p-1} \gamma_i \Delta y_{t-i} + \varepsilon_t. \quad (5.15)$$

Dickey and Fuller (1981) also provided many theories for testing for a unit root in univariate time series. They demonstrated that an F-type likelihood ratio test statistic Φ_1 (for more information, see Dickey and Fuller (1981)) has greater power over the other rival test statistics. This test is founded on a regression, such as in (5.16), where the initial value y_0 is treated as fixed.

$$y_t = \mu + \pi_1 y_{t-1} + \pi_2 y_{t-2} + \varepsilon_t, \quad t = 1, 2, \dots, T. \quad (5.16)$$

When one rewrites (5.16) in a different way, we get:

$$\Delta y_t = \mu + \rho y_{t-1} + \gamma \Delta y_{t-1} + \varepsilon_t. \quad (5.17)$$

where $\rho = (\pi_1 + \pi_2 - 1)$, $\gamma = -\pi_2$ and μ is the mean of the process. The process in (5.16) is said to contain a unit root if it is difference stationary (DS) and (5.17) is not, i.e., if $\rho = 0 \Rightarrow \pi_1 + \pi_2 = 1$. Thus, this method tests null hypothesis $H_0 : \mu = 0$ and $\rho = 0$ against the alternative hypothesis $H_a : \mu \neq 0$ and $\rho \neq 0$. In other words, it compares if the series is $I(1)$ with nonzero drift against $I(0)$ with zero drift. If the calculated value of test statistic Φ_1 is larger than a critical value, then, the null hypothesis is rejected.

5.6.2 Phillips-Perron Test (*PP* Test)

Phillips and Perron (1988) developed a number of unit root tests that have become popular in the analysis of financial time series. The Phillips-Perron (*PP*) unit root tests differ from the ADF tests mainly in how they deal with serial correlation and heteroscedasticity in the errors. In particular, where the ADF tests use a parametric autoregression to approximate the ARMA structure of the errors in the test regression, the *PP* tests ignore any serial correlation in the test regression. The test regression for the *PP* tests is:

$$\Delta y_t = \beta' \mathbf{D}_t + \pi y_{t-1} + u_t \quad (5.18)$$

where u_t is $I(0)$ and may be heteroscedastic. The *PP* tests correct for any serial correlation and heteroscedasticity in the errors u_t of the test regression by directly modifying the test statistics $t_{\pi=0}$ and $T\hat{\pi}$. These modified statistics, denoted Z_t and Z_π , are given by

$$\begin{aligned} Z_t &= \left(\frac{\hat{\sigma}^2}{\hat{\lambda}^2} \right)^{1/2} \cdot t_{\pi=0} - \frac{1}{2} \left(\frac{\hat{\lambda}^2 - \hat{\sigma}^2}{\hat{\lambda}^2} \right) \cdot \left(\frac{T \cdot SE(\hat{\pi})}{\hat{\sigma}^2} \right) \\ Z_\pi &= T\hat{\pi} - \frac{1}{2} \frac{T^2 \cdot SE(\hat{\pi})}{\hat{\sigma}^2} (\hat{\lambda}^2 - \hat{\sigma}^2) \end{aligned} \quad (5.19)$$

The terms $\hat{\sigma}^2$ and $\hat{\lambda}^2$ are consistent estimates of the variance parameters

$$\begin{aligned}\sigma^2 &= \lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T E[u_t^2] \\ \lambda^2 &= \lim_{T \rightarrow \infty} \sum_{t=1}^T E[T^{-1} S_T^2]\end{aligned}\tag{5.20}$$

where $S_T = \sum_{t=1}^T u_t$. The sample variance of the least squares residual \hat{u}_t is a consistent estimate of σ^2 , and the Newey-West long-run variance estimate of u_t using \hat{u}_t is a consistent estimate of λ^2 .

Under the null hypothesis that $\pi = 0$, the *PP* Z_t and Z_π statistics have the same asymptotic distributions as the ADF t-statistic and normalised bias statistics. One advantage of the *PP* tests over the ADF tests is that the *PP* tests are robust to general forms of heteroscedasticity in the error term u_t . Another advantage is that the user does not have to specify a lag length for the test regression.

If unit root testing is generated by the heteroscedastic and autocorrelated non-systematic component, it is difficult to estimate lag p in the regression model. To describe the autocorrelation structure of the producing process, Phillips and Perron (1988) used the standard DF test with non-parametrically adjusted test statistics instead of employing proper autocorrelation models. This test is based on the model equations (5.10) and (5.13), and with the exception of the last model, the linear trend is substituted by a centered time variable (see equation (5.18)).

5.6.3 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

The Kwiatkowski Phillips Schmidt Shin (KPSS) test examines if a time series fluctuates around a linear trend or mean or is non-stationary or stationary due to a unit root. This test is also known as the stationarity test (Sephton, 1995). Against a unit root alternative, the null hypothesis is that a process is trend stationary. Dickey and Fuller (1979b) presented a test statistic based on the Lagrange multiplier (LM) equation (5.21)

$$\begin{aligned}y_t &= \theta t + w_t + \varepsilon_t, \\ w_t &= w_{t-1} + u_t.\end{aligned}\tag{5.21}$$

where w_t is a random walk with fixed initial value w_0 , $u_t \stackrel{\text{iid}}{\sim} N(0, \sigma_u^2)$ and $\varepsilon_t \stackrel{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$.

The KPSS test is used to test the null hypothesis that the time series $H_0 : \sigma_u^2 = 0$ versus $H_a : \sigma_u^2 > 0$, because ε_t is $I(0)$ under the stationarity assumption (Kwiatkowski and colleagues, 1992). The null hypothesis states that the time series y_t is integrated of order one, $I(1)$, as tested by all of the above tests. The KPSS test describes the opposite case, namely testing the null hypothesis that the time series y_t is $I(0)$ (Dickey and Fuller, 1979b).

The linear regression is used in the KPSS test. The regression equation divides a series into three parts: a random walk (r_t), a deterministic trend (β_t), and a stationary error (ε_t), with the regression equation:

$$x_t = r_t + \beta t + \varepsilon_1. \quad (5.22)$$

In cases where the data is stationary, the series will be stationary or the intercept will be around a fixed level (Wang, 2006).

The test finds the equation using ordinary least squares (OLS), which varies somewhat contingent on whether the researcher wants to test for level or trend stationarity (Kočenda and Černý, 2015). Before conducting the KPSS test, data is usually log-transformed to convert any exponential into linear trends.

5.7 Autoregressive Integrated Moving Average (ARIMA) Process

ARIMA models come in a wide range of shapes and sizes (Box and Jenkins, 1968). ARIMA (p, d, q) , where p refers to the number of autoregressive terms, d refers to the number of differences, and q refers to the number of moving average terms, is an overall non-seasonal model. A white noise model is characterized as ARIMA $(0, 0, 0)$ since there is no AR part because it is independent of y_{t-1} , there is no differencing, and there is no MA part because y_t is independent of e_{t-1} . Briefly, the ARIMA univariate analysis models consist of 3 sub-processes: model identification, parameter estimation and model diagnosis. By repeating these three steps, the optimal prediction model is screened out

(Chadsuthi et al., 2012).

If y_t is non-stationary, for example, we can take a first-difference of y_t to make it stationary.

$$\begin{aligned}\Delta y_t &= y_t - y_{t-1} \text{ (} d = 1 \text{ implies one time differencing)}, \\ \Delta y_t &= c + \alpha_1 \Delta y_{t-1} + \alpha_2 \Delta y_{t-2} + \cdots + \alpha_p \Delta y_{t-p} + \theta_1 u_{t-1} + \\ &\quad \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q} + u_t.\end{aligned}\tag{5.23}$$

refers to an ARIMA $(p, 1, q)$ model.

Since there is no AR or MA component and only one difference, a random walk model is classed as ARIMA $(0, 1, 0)$.

The following steps are used to fit the model:

- **FIRSTLY** The Augmented Dickey-Fuller (ADF) test and the disease sequence diagram are used to determine the initial sequence's stationarity. In cases where the sequence is non-stationary, the first-order seasonal difference ($D = 1$) and the first-order ordinary difference ($d = 1$) and are used to stabilise it and remove the seasonality and trend. The stationary series is investigated further.
- **SECONDLY** The study looks at the partial autocorrelation function (PACF) and the autocorrelation function (ACF) graphs to calculate the model parameters, p and q . These parameters are then estimated using the maximum likelihood estimation (MLE) approach. Next, the residuals and parameters of the established ARIMA model are examined to determine its acceptability, and the Ljung-Box (Q) test (Ljung and Box, 1978) is used to determine if the residuals of the model are white noise (Davies et al., 1977).
- **FINALLY** If multiple models satisfy the condition of having significant parameters and the residual sequence of the model is white noise, the optimal univariate model can be chosen using the Schwarz Bayesian information criterion (SBC), Akaike information criterion (AIC), root mean square error (RMSE) and mean absolute percentage error (MAPE) indicators.

5.8 GARCH Model

To overcome the limitations of the standard autoregressive conditionally heteroscedastic (ARCH) model, Bollerslev (1986) presented a generalised ARCH (GARCH) (Engle, 1982). Both a longer memory and a more flexible lag structure were possible with the GARCH model. The conditional variance in the ARCH process is expressed as a linear mapping of prior sample variance alone, but the GARCH process also allows for lagged conditional variances to be included in the model. Bollerslev (1986) and Engle (1982) could not distinguish between how positive and negative values affected the variance of the return in their GARCH and ARCH models.

For GARCH models, the general equation is:

$$\begin{aligned} Y_t &= \mu + X_t \varepsilon_t, \\ X_t &= \sigma_t \varepsilon_t, \\ \sigma_t^2 &= \alpha_0 + \alpha_1 X_{t-1}^2 \dots \alpha_p X_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \dots + \beta_q \sigma_{t-q}^2. \end{aligned} \tag{5.24}$$

In this case, Y_t is the series return, and μ is the average series; ε_t are separate and identical distribution chains that trace the standard normal distribution with an average of 0 and variance of 1. Equation (5.24) can be rewritten as:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^p \alpha_i X_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2. \tag{5.25}$$

where the model parameters are represented by $\alpha_0 \geq 0, \alpha_i \geq 0, \beta_j \geq 0$ for $i > 0, j > 0$. When $p = 1$ and $q = 1$, the GARCH model is expressed as:

$$\begin{aligned} Y_t &= \mu + X_t, \\ X_t &= \sigma_t \varepsilon_t, \\ \sigma_t^2 &= \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \end{aligned} \tag{5.26}$$

This model can be statistically compared using the Hannan-Quinn information criterion (HQC) and Akaike information criterion (AIC) to see if it is suitable for forecasting

purposes (Pereira et al., 2004).

5.9 Vector Autoregression (VAR)

Multivariate time series (MTS) models come in a variety of shapes and sizes. The decision between linear and non-linear models (Casdagli, 1992) comes first, followed by the precise type of model within these two classifications. Many non-linear models have been specifically constructed and tailored for the issue domain they are used to, and deciding which model to apply, particularly in the financial realm, can be difficult. One of the multivariate linear time series models will be used to model the dataset under discussion in this study. The Vector Moving Average (VMA) process, the Vector AutoRegressive (VAR) process, and the Vector AutoRegressive Moving Average (VARMA) process are the three basic linear models (Lütkepohl, 2013).

Vector autoregression (VAR) was developed by Sims (1980) and it was established as a method for macroeconomists to use when describing the joint dynamic behaviour of a set of variables deprived of the need for severe constraints, such as those required for detecting underlying structural factors. Since then it has become a widely used time series modeling technique. Although not always required, strong identification assumptions are needed for some of the most practical applications of the estimates. Examples of such include computing variance decompositions or impulse-response functions (IRFs). A common constraint is defined as an assumption about a pair of variables' dynamic correlation. For instance, x affects y only with m variables in a VAR system as well as p lags and an error term. While exogenous variables like time trends or seasonal dummies can also be included in a VAR, it is more important to understand the basics first. An order- p VAR consists of two equations with two variables, x and y :

$$\begin{aligned} y_t &= \beta_{y0} + \beta_{yy1}y_{t-1} + \dots + \beta_{yyp}y_{t-p} + \beta_{yx1}x_{t-1} + \dots + \beta_{yxp}x_{t-p} + v_t^y, \\ x_t &= \beta_{x0} + \beta_{xy1}y_{t-1} + \dots + \beta_{xyp}y_{t-p} + \beta_{xx1}x_{t-1} + \dots + \beta_{xrp}x_{t-p} + v_t^x. \end{aligned} \tag{5.27}$$

The coefficient of y in the equation for x at lag p is denoted by β_{xyp} using the subscript convention. If another variable z is added to the system, a third equation for z_t will be created, and phrases using p delayed or lagged values of z , such as β_{xzp} , will be included to each of the three equations, on the right-hand side.

The fact that none of the variables currently exist on the right-hand side of the equations is a significant aspect of equations (5.27). This suggests that the regressors in (5.27) are not strongly exogenous and that, provided that the variables are ergodic and stationary, OLS can generate asymptotically suitable estimators, albeit this is not always the case. Variables that have been determined to be exogenous, such as seasonal dummy variables, can be easily included on the right-hand side of the VAR equations, and without the need for extra equations to describe them. Such exogenous factors will not be included in our instances.

A vector error-correction (VEC) model is used when the variables of a VAR are cointegrated (Engle and Granger, 1987). The following is an example of a VEC with two variables:

$$\begin{aligned}
 \Delta y_t &= \beta_{y0} + \beta_{y1}\Delta y_{t-1} + \dots + \beta_{yp}\Delta y_{t-p} + \gamma_{y1}\Delta x_{t-1} + \dots + \gamma_{yp}\Delta x_{t-p} \\
 &\quad - \lambda_y (y_{t-1} - \alpha_0 - \alpha_1 x_{t-1}) + v_t^y, \\
 \Delta x_t &= \beta_{x0} + \beta_{x1}\Delta y_{t-1} + \dots + \beta_{xp}\Delta y_{t-p} + \gamma_{x1}\Delta x_{t-1} + \dots + \gamma_{xp}\Delta x_{t-p} \\
 &\quad - \lambda_x (y_{t-1} - \alpha_0 - \alpha_1 x_{t-1}) + v_t^x.
 \end{aligned} \tag{5.28}$$

where the long-run cointegrating connection between the two variables is $y_t = \alpha_0 + \alpha_1 x_t$, and the error-correction parameters are λ_y and λ_x , which quantify how y and x react to departures from long-run equilibrium.

5.9.1 Forecasting using VAR Model

The equations' structure in (5.27) is intended to represent how the values of the variables in period t are connected to previous values. As a result, the VAR is well-suited to

estimating the future courses of x and y based on their prior histories. Assume the study has a sample of x and y observations that ends in period T , and we want to forecast their values in $T + 1$, $T + 2$, and so on. Assume that $p = 1$, which means that there is just one lagged value on the right-hand side. Our VAR for period $T + 1$ is

$$\begin{aligned} y_{T+1} &= \beta_{y0} + \beta_{yy1}y_T + \beta_{yx1}x_T + v_{T+1}^y, \\ x_{T+1} &= \beta_{x0} + \beta_{xy1}y_T + \beta_{xx1}x_T + v_{T+1}^x. \end{aligned} \tag{5.29}$$

Taking the expectation as a function of the sample's relevant information (x_T and y_T) yields

$$\begin{aligned} E(y_{T+1} | x_T, y_T) &= \beta_{y0} + \beta_{yy1}y_T + \beta_{yx1}x_T + E(v_{T+1}^y | x_T, y_T), \\ E(x_{T+1} | x_T, y_T) &= \beta_{x0} + \beta_{xy1}y_T + \beta_{xx1}x_T + E(v_{T+1}^x | x_T, y_T). \end{aligned} \tag{5.30}$$

In order for OLS to estimate the coefficients consistently, the conditional expectation of the VAR error components on the right-hand side must be zero. The serial correlation features of the v terms will determine whether or not this assumption is correct; as we've seen, serially correlated mistakes and lagged dependent variables like those found in the VAR can be a toxic mix.

As a result, we must ensure that $E(v_t^j | v_{t-1}^x, v_{t-1}^y) = 0$. We assume that our VAR system has a long enough lag for the error term to be non-serially correlated, and that the conditional expectation of the error term for all periods following T is zero. This means that the right-hand side of each equation in (5.30) has a zero final term, thus the projections are

$$\begin{aligned} Pr(y_{T+1} | x_T, y_T) &\equiv \hat{y}_{T+1|T} = \hat{\beta}_{y0} + \hat{\beta}_{yy1}y_T + \hat{\beta}_{yx1}x_T, \\ Pr(x_{T+1} | x_T, y_T) &\equiv \hat{x}_{T+1|T} = \hat{\beta}_{x0} + \hat{\beta}_{xy1}y_T + \hat{\beta}_{xx1}x_T. \end{aligned} \tag{5.31}$$

The forecast error in the predictions in equation (5.31) will come from two sources: the unpredictable error term at time $T + 1$ and the errors we make in estimating the β coefficients. Formally,

$$\begin{aligned}
y_{T+1} - \hat{y}_{T+1|T} &= (\beta_{y0} - \hat{\beta}_{y0}) + (\beta_{yy1} - \hat{\beta}_{yy1}) y_T + (\beta_{yx1} - \hat{\beta}_{yx1}) x_T + v_{T+1}^y, \\
x_{T+1} - \hat{x}_{T+1|T} &= (\beta_{x0} - \hat{\beta}_{x0}) + (\beta_{xy1} - \hat{\beta}_{xy1}) y_T + (\beta_{xx1} - \hat{\beta}_{xx1}) x_T + v_{T+1}^x.
\end{aligned} \tag{5.32}$$

If our estimates of the β coefficients are constant and there is no sequential correlation in v , then the expectation of the forecast error is asymptotically zero. The variance of the forecast error is

$$\begin{aligned}
\text{var}(y_{T+1} - \hat{y}_{T+1|T}) &= \text{var}(\hat{\beta}_{y0}) + \text{var}(\hat{\beta}_{yy1}) y_T^2 + \text{var}(\hat{\beta}_{yx1}) x_T^2 \\
&\quad + 2 \text{cov}(\hat{\beta}_{y0}, \hat{\beta}_{yy1}) y_T + 2 \text{cov}(\hat{\beta}_{y0}, \hat{\beta}_{yx1}) x_T + 2 \text{cov}(\hat{\beta}_{yy1}, \hat{\beta}_{yx1}) x_T y_T \\
&\quad + \text{var}(v_{T+1}^y), \\
\text{var}(x_{T+1} - \hat{x}_{T+1|T}) &= \text{var}(\hat{\beta}_{x0}) + \text{var}(\hat{\beta}_{xy1}) y_T^2 + \text{var}(\hat{\beta}_{xx1}) x_T^2 \\
&\quad + 2 \text{cov}(\hat{\beta}_{x0}, \hat{\beta}_{xy1}) y_T + 2 \text{cov}(\hat{\beta}_{x0}, \hat{\beta}_{xx1}) x_T + 2 \text{cov}(\hat{\beta}_{xy1}, \hat{\beta}_{xx1}) x_T y_T \\
&\quad + \text{var}(v_{T+1}^x).
\end{aligned} \tag{5.33}$$

With the exception of the last item, all of the terms in this formula converge to zero as our consistent estimates of the β coefficients converge to the true values (as T increases). As a result, the error in estimating the coefficients is frequently overlooked when calculating the variance of the forecast error, resulting in

$$\begin{aligned}
\text{var}(y_{T+1} - \hat{y}_{T+1|T}) &\approx \text{var}(v_{T+1}^y) \equiv \sigma_{v,y}^2, \\
\text{var}(x_{T+1} - \hat{x}_{T+1|T}) &\approx \text{var}(v_{T+1}^x) \equiv \sigma_{v,x}^2.
\end{aligned} \tag{5.34}$$

The ability to employ the VAR recursively to extend forecasts into the future is one of its most useful features. For the time span $T + 2$,

$$\begin{aligned}
E(y_{T+2} | x_{T+1}, y_{T+1}) &= \beta_{y0} + \beta_{yy1} y_{T+1} + \beta_{yx1} x_{T+1}, \\
E(x_{T+2} | x_{T+1}, y_{T+1}) &= \beta_{x0} + \beta_{xy1} y_{T+1} + \beta_{xx1} x_{T+1}.
\end{aligned} \tag{5.35}$$

So, by recursive expectations, we have:

$$\begin{aligned}
E(y_{T+2} | x_T, y_T) &= \beta_{y0} + \beta_{yy1}E(y_{T+1} | x_T, y_T) + \beta_{yx1}E(x_{T+1} | x_T, y_T) \\
&= \beta_{y0} + \beta_{yy1}(\beta_{y0} + \beta_{yy1}y_T + \beta_{yx1}x_T) + \beta_{yx1}(\beta_{x0} + \beta_{xy1}y_T + \beta_{xx1}x_T), \\
E(x_{T+2} | x_T, y_T) &= \beta_{x0} + \beta_{xy1}E(y_{T+1} | x_T, y_T) + \beta_{xx1}E(x_{T+1} | x_T, y_T) \\
&= \beta_{x0} + \beta_{xy1}(\beta_{y0} + \beta_{yy1}y_T + \beta_{yx1}x_T) + \beta_{xx1}(\beta_{x0} + \beta_{xy1}y_T + \beta_{xx1}x_T).
\end{aligned} \tag{5.36}$$

To obtain the necessary projections (forecasts), coefficient estimates are substituted once again.

$$\begin{aligned}
Pr(y_{T+2} | x_T, y_T) &\equiv \hat{y}_{T+2|T} = \hat{\beta}_{y0} + \hat{\beta}_{yy1}\hat{y}_{T+1|T} + \hat{\beta}_{yx1}\hat{x}_{T+1|T}, \\
Pr(x_{T+2} | x_T, y_T) &\equiv \hat{x}_{T+2|T} = \hat{\beta}_{x0} + \hat{\beta}_{xy1}\hat{y}_{T+1|T} + \hat{\beta}_{xx1}\hat{x}_{T+1|T}.
\end{aligned} \tag{5.37}$$

If we once again ignore error in estimating the coefficients, then the two-period-ahead forecast error in (5.9.1) is

$$\begin{aligned}
y_{T+2} - \hat{y}_{T+2|T} &\approx \beta_{yy1}(y_{T+1} - y_{T+1|T}) + \beta_{yx1}(x_{T+1} - x_{T+1|T}) + v_{T+2}^y \\
&\approx \beta_{yy1}v_{T+1}^y + \beta_{yx1}v_{T+1}^x + v_{T+2}^y, \\
x_{T+2} - \hat{x}_{T+2|T} &\approx \beta_{xy1}(y_{T+1} - y_{T+1|T}) + \beta_{xx1}(x_{T+1} - x_{T+1|T}) + v_{T+2}^x \\
&\approx \beta_{xy1}v_{T+1}^y + \beta_{xx1}v_{T+1}^x + v_{T+2}^x.
\end{aligned} \tag{5.38}$$

The error terms for period $T+1$ will be correlated across equations in general. Therefore the variance of the two-period forecast will be roughly:

$$\begin{aligned}
\text{var}(y_{T+2} - \hat{y}_{T+2|T}) &\approx \beta_{yy1}^2\sigma_{v,y}^2 + \beta_{yx1}^2\sigma_{v,x}^2 + 2\beta_{yy1}\beta_{yx1}\sigma_{v,xy} + \sigma_{v,y}^2 \\
&= (1 + \beta_{yy1}^2)\sigma_{v,y}^2 + \beta_{yx1}^2\sigma_{v,x}^2 + 2\beta_{yy1}\beta_{yx1}\sigma_{v,xy}, \\
\text{var}(x_{T+2} - \hat{x}_{T+2|T}) &\approx \beta_{xy1}^2\sigma_{v,y}^2 + \beta_{xx1}^2\sigma_{v,x}^2 + 2\beta_{xy1}\beta_{xx1}\sigma_{v,xy} + \sigma_{v,x}^2 \\
&= \beta_{xy1}^2\sigma_{v,y}^2 + (1 + \beta_{xx1}^2)\sigma_{v,x}^2 + 2\beta_{xy1}\beta_{xx1}\sigma_{v,xy}.
\end{aligned} \tag{5.39}$$

Since the errors made in forecasting period $T+1$ spread into errors in the forecast for $T+2$, the two-period-error forecast error has a higher variance than the one-period-ahead error. The variation grows as our prediction horizon grows, indicating our inability to foresee far into the future, even if we had correct estimations of the coefficients (as we have optimistically assumed here).

When larger forecast horizons are considered, the computations in equation (5.39) get more complicated. In both (5.9.1) and (5.39), including more than one lag on the right-hand side or more than two variables in the VAR more than significantly increases the number of terms. These processes have been automated for us thanks to current statistical software such as Gretl, RStudio, E-Views, Python and STATA.

5.9.2 Lag Length Selection Using Information Criteria

Economic theory can occasionally drive the choice of lag lengths in autoregressive distributed lag (ADL) and AR models. There are, however, statistical approaches for determining how many lags should be used as regressors. Overall, too many lags increase the standard errors of coefficient estimates, implying a rise in forecast error, whereas removing lags that should be involved in the model might lead to estimation bias.

There are two methods for determining the order of an AR model:

- **THE F-TEST APPROACH** Calculate an $AR(p)$ model and examine the importance of the longest lag(s). If the test fails, remove the lag(s) in question from the model. This method tends to yield models with excessive order: we always run the risk of rejecting a true null hypothesis in a significance test!
- **RELYING ON AN INFORMATION CRITERION** We can utilise the criteria's minimal values to determine the length of the lag to be chosen. The following are some examples of regularly used criteria:

- Final Prediction Error (FPE) (Akaike, 1969):

$$FPE(p) = \left(\frac{T+p^*}{T-p^*} \right)^n |\hat{S}(p)|; \quad (5.40)$$

- The Akaike information criterion (AIC) (Akaike, 1973):

$$AIC(p) = \ln(|\hat{S}(p)|) + \frac{2n^2p}{T}; \quad (5.41)$$

- The Schwarz Bayesian Criterion (SBC) (Schwarz, 1978):

$$BIC(p) = \ln(|\hat{S}(p)|) + \frac{\ln T}{T}pn^2; \quad (5.42)$$

- Hannan-Quinn Criterion (HQC) (Hannan and Quinn, 1979):

$$HQC(p) = \ln(|\hat{S}(p)|) + \frac{2\ln(\ln(T))}{T}pn^2. \quad (5.43)$$

where $\hat{S}(p) = \frac{1}{T} \sum_{t=1}^T \hat{e}_t (\hat{e}_t)$ represents the residuals' estimated covariance matrix from the model $\text{VAR}(p)$, and p represents the order for a given vector time series, T represents the number of observations, and n represents the length of the model VAR, where, p^* is the total number of parameters in each.

The observed time series were minimally Gaussianized and deseasonalised before being included in any of the models we presented. It is conceivable to test auto-regression order from 1 to 20 days or more using the function "VARselect" in the R package "vars" (Pfaff et al., 2008) to determine the optimal values based on the information criteria; for instance, in our reference study (the association between rheumatology disease activity score DAS28 and air pollution), the auto-regression of order seven was chosen based on the outcomes of Akaike Information Criterion (AIC) and Bayes information criterion (BIC or SBC). We also used some GRETL-provided criteria, such as the Finite Prediction Error (FPE), Akaike Information Criterion (AIC), Hannan and Quin Criterion (HQC) and Schwarz Bayesian criterion (BIC or SBC).

5.10 Johansen and Juselius Cointegration Test

Because the variables to be utilised are unlikely to be stationary, Granger and Newbold (1974) point out that using OLS on the level variables will generate erroneous results.

Engle and Granger (1987) proposed a two-step method based on a unit root test of the residuals and OLS estimation of the long-run equation. When the residual series is proven to be stationary, it is used as an error correction term in the (differenced) short-run specification. The error correction (EC) term indicates how quickly the long run equation is adjusted to equilibrium. Even while it is popular, it has flaws. There may be more than one cointegrating vector in a model if there are more than two variables. These different cointegrating vectors cannot be detected using the single-equation approach. Even if there is just one cointegrating vector, the univariate technique is inefficient if not all variables on the cointegrating vector's right-hand side are weakly exogenous. The vector error correction (VEC) is one of the numerous cases of the VAR method that is based on the variables that are stationary in their differences (i.e., $I(1)$). The VEC method can also take into account any cointegrating relationships among the variables and that is the purpose why we use it in this thesis. The vector error correction (VEC) was proposed and developed by Johansen (1991, 1988) and Johansen (1992). The Trace test and the Maximum Eigenvalue test are two tests used by Johansen procedures Johansen and Juselius (1990) to identify the number of cointegration vectors. For $r = 0, 1, 2 \dots n - 1$, the Maximum Eigenvalue statistic compares the null hypothesis of r cointegrating relations to the alternative of $r + 1$ cointegrating relations. These test statistics are calculated as follows:

$$LR_{\max}(r/n + 1) = -T * \log(1 - \hat{\lambda}), \quad (5.44)$$

where T is the sample size and λ is the maximum eigenvalue. For $r = 0, 1, 2 \dots n - 1$, trace statistics compare the null hypothesis of r co-integrating relations to the alternative of n co-integrating relations, where n is the number of variables in the system. The following formula is used to calculate its equation:

$$LR_{tr}(r/n) = -T * \sum_{i=r+1}^n \log(1 - \hat{\lambda}_i). \quad (5.45)$$

In rare circumstances, the findings of the Trace and Maximum Eigenvalue statistics

may differ, and Alexander (2001) suggests that the outcomes of the trace test should be favoured.

5.11 Granger Causality Test

Because regression analysis simply considers one variable's dependency on other variables, it does not always imply causality or have predictive value (MacNally, 2000). In contrast, Granger causality tests are used to analyse the causal relationship between variables. To determine the presence and direction of causation between the variables under examination, this study uses the widely utilised Granger causality test. The link between variables is determined by the direction of causality. There could be one-way causation, two-way causality, or no causality between the variables. According to the Granger causality test, if a variable X causes variable Y , the mean square error of a forecast of Y based on past values of X is lower than that of a forecast based on simply previous values of Y . To perform a Granger causality test, we start by assuming that all variables are stationary. If the original variables have unit roots, we can conclude that differences have been made to account for the original variables' changes in the model (which do not have unit roots). We started with an Autoregressive Distributed Lag Model $ADL(p, q)$ model for Y as the dependent variable for investigating Granger causation between X and Y . The model $ADL(p, q)$ presupposes that a time series Y_t may be characterised by a linear function of q lags and p lagged values of another time series X_t :

$$\begin{aligned}
 Y_t = & \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \\
 & + \delta_1 X_{t-1} + \delta_2 X_{t-2} + \cdots + \delta_q X_{t-q} + u_t.
 \end{aligned}
 \tag{5.46}$$

is an autoregressive distributed lag model with q lags of X_t and p lags of Y_t and where

$$E(u_t | Y_{t-1}, Y_{t-2}, \dots, X_{t-1}, X_{t-2}, \dots) = 0.
 \tag{5.47}$$

We utilised this model to detect if X Granger caused Y . After that, we looked at

causation in the opposite manner, reversing the roles of X and Y in the ADL. X became the dependent variable in particular. The two equations can be written as follows:

$$\begin{aligned} Y_t &= \alpha_1 + \delta_{1t} + \phi_{11}Y_{t-1} + \dots + \phi_{1p}Y_{t-p} + \beta_{11}X_{t-1} + \dots + \beta_{1q}X_{t-q} + \epsilon_{1t}, \\ X_t &= \alpha_2 + \delta_{2t} + \phi_{21}Y_{t-1} + \dots + \phi_{2p}Y_{t-p} + \beta_{21}X_{t-1} + \dots + \beta_{2q}X_{t-q} + \epsilon_{2t}. \end{aligned} \quad (5.48)$$

The first equation determines if X Granger causes Y , while the second determines whether Y Granger causes X . The coefficients now contain subscripts showing which equation they belong to. Subscripts have been added to the error terms to indicate that they will be different in the two equations.

5.12 Cointegration: Empirical Background

The distinction between non-stationary and stationary time series, as well as weak and rigorous stationarity, is critical. This is important for stock market cointegration analysis since we anticipate stock prices to be non-stationary (Richards, 1995). If the probability distribution of a time series' values does not vary over time, it is said to be strictly stationary (Brooks, 2019):

$$f(y_t, y_{t+1}, \dots, y_T) = f(y_{t+k}, y_{t+1+k}, \dots, y_{T+k}). \quad (5.49)$$

Strict stationarity requires that all higher-order moments, such as mean and variance, remain constant. In practice, however, strictly stationary time series are uncommon. As a result, in our next investigation, we will concentrate on weakly stationary processes. Weakly stationary processes might be considered stationary if their assumptions and conditions are met. When the mean, variance, and autocovariance of a time series remain consistent across time, it is considered weakly stationary (Enders, 2008).

Non-stationary time series, on the other hand, have features that change with time. At different time points in this type of time series, the variance and mean have different values. As the sample size grows larger, the variance will increase (Korkas and Pry-

zlewiczV, 2017). There are various reasons why distinguishing between non-stationary and stationary series is significant. We'll demonstrate this with the help of a simple autoregressive (AR) process:

$$y_t = \mu + \rho y_{t-1} + u_t. \quad (5.50)$$

where the present value of variable y is determined by the constant term μ , the variable y 's value from the previous period $t-1$, and an error term u_t . We're particularly interested in the value of ρ since it indicates whether the process is stationary or non-stationary. There are three different scenarios that could happen, or three different values that could be used of ρ (Brooks, 2019):

1. $|\rho| < 1$; a shock to the system in the present period t is transient; it will fade away with time, and this series is stationary because its variance, mean, and autocovariance are all constant. In the long run, a stationary time series will return to its mean value ("Mean reversion").
2. $\rho = 1$; A shock in time will not fade away with time, but will remain everlasting. Over time, its variance will approach infinity. This time series is non-stationary, or the unit root case, since the variable y contains a unit root.
3. $\rho > 1$; a shock in the timeline, since this type of time series is likewise non-stationary, will explode with time. Over time, there is no mean reversion to its true value.

Non-stationary variables are integrated of order d , where $d \geq 1$: $y_t \sim I(d)$, while stationary variables are integrated of order 0, denoted $y_t \sim I(0)$. Only the values $d = 0$ and $d = 1$ will be considered in the remainder of the thesis.

By taking the difference one or more times, non-stationary variables can be turned into stationary variables. If a time series has one unit root (order one integration), then subtracting the difference once makes the time series variable stationary. Similarly,

subtracting the difference d -times from a non-stationary variable with d unit roots (integrated of order d) converts the variable to a stationary variable (Enders, 2008).

The work of Engle and Granger laid the foundation for the notion of cointegration (1987). Two variables are cointegrated if they have the same long-term stochastic trend. When two integrated variables are combined, the higher of the two integration orders is always used. In time series, either zero or one is the most common order of integration (Brooks, 2019):

1. if $y_t \sim I(0)$, and $x_t \sim I(0)$, then their combination $(ax_t + by_t)$ will also be $I(0)$,
2. if $y_t \sim I(0)$, and $x_t \sim I(1)$, then their combination $(ax_t + by_t)$ will now be $I(1)$, because $I(1)$ is a higher order of integration and dominates the lower order of integration $I(0)$,
3. if $y_t \sim I(1)$, and $x_t \sim I(1)$, then their combination $(ax_t + by_t)$ will also be $I(1)$, in the general case.

Cointegration between non-stationary $I(1)$ variables exists if there is a linear combination of them that is stationary, $I(0)$.

Two $I(1)$ non-stationary variables, y_t and x_t , are included in the following regression model:

$$y_t = \mu + \beta x_t + u_t. \tag{5.51}$$

These two variables are cointegrated if the OLS estimate β makes the linear combination of y_t and x_t stationary. The error term between them becomes stationary with time:

$$u_t = y_t - \beta x_t. \tag{5.52}$$

Two variables must be integrated with the same order for them to be cointegrated. They cannot be cointegrated, for example, if one order is $I(0)$ and the other is $I(1)$. If two variables are integrated with different orders, cointegration will not exist since the highest order of integration of the two variables will dominate.

5.13 Cointegration: Model and Notation

Assume that the k -dimensional vector process \mathbf{z}_t presented in Section 5.3 is an order 1 integrated process. The VAR(p) model in equation (5.5) can be given several parameterizations for such a process without placing any binding limits on the model parameters, i.e. without modifying the likelihood function's value. In differenced form, the equivalent of equation (5.5) can be written as:

$$\Delta \mathbf{z}_t = \Pi \mathbf{z}_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta \mathbf{z}_{t-j} + \varepsilon_t, \quad t = 1, 2, \dots, T. \quad (5.53)$$

where $\Delta = 1 - L$, $\Gamma_j = -\sum_{i=j+1}^p \Pi_i$ and $\Pi = \sum_{i=1}^p \Pi_i - \mathbf{I}_k$. Here the long-run matrix $\Pi = -\Pi^*(1)$ for $\Pi^*(1)$ discussed in Section 5.4, which should be singular if \mathbf{Z}_t is a vector integrated time series.

Note that inclusion of $\Pi \mathbf{z}_{t-1}$ in equation (5.53) can raise a question of how to handle the non-stationarity problem while setting $\Pi \mathbf{z}_{t-1} = \mathbf{0}$ will leave a model with only the short-run dynamics (long-run information will be lost). Thus equation (5.53) makes sense only if $\Pi \mathbf{z}_{t-1}$ defines stationary linear combinations of the I(1) variables, in which case the reparametrized model is in vector error correction model (VECM) form. According to Engle and Granger (1987) VECM implies cointegration.

A $k \times 1$ vector time series \mathbf{Z}_t is assumed to be cointegrated (Hamilton, 1994) if each of the series is I(1), or non-stationary with a unit root, while some (at least one) linear combinations of the series $\beta_i' \mathbf{z}_{t-1}$ are stationary, or I(1), for any nonzero $k \times 1$ vector β_i called the cointegrating vector. The cointegrating rank (Johansen, 2000) is defined as the number r of linearly independent cointegrating vectors, and the cointegration space is defined as the space spanned by the cointegrating vectors.

As a result, the cointegration hypothesis may be expressed within the VAR(p) model as a reduced rank limitation on the long-run matrix Π , such that $\Pi = \alpha \beta'$, for α and β with $p \times r$ matrices, and with β_i ($i = 1, 2, \dots, r$) being the i^{th} column of the cointegrating matrix β . The loading or adjustment matrix (α) is a matrix whose members determine the adjustment of change to the long-run equilibrium (Juselius, 2006).

5.14 Cointegration Analysis

If the analytical variables are non-stationary and integrated in order d , $I(d)$, we can difference d times to make their differences stationary. However, the factors may have a cointegrating relationship. If two non-stationary variables are integrated in the same order d , $I(d)$, a linear combination of these two variables may be stationary, which is known as cointegration. If the two variables have a cointegrating relationship, differencing will result in a loss of information. Let's say we've got the following equation:

$$x_{1t} = \beta_1 + \beta_2 x_{2t} + \dots + \beta_n x_{nt} + e_t. \quad (5.54)$$

We get the following equation when we solve for the error term:

$$e_t = x_{1t} - \beta_1 - \beta_2 x_{2t} - \dots - \beta_n x_{nt}. \quad (5.55)$$

In light of this, Engle and Granger (1987) proposed the notion of cointegration with a collection of long-run equilibrium variables.

$$\beta_1 x_{1t} - \beta_2 x_{2t} - \dots - \beta_n x_{nt} = 0. \quad (5.56)$$

In this case $\beta = (\beta_1, \dots, \beta_n)$ is the cointegrating vector. We can formulate the following equation because the divergence from long-run equilibrium is e_t :

$$\beta_1 x_{1t} - \beta_2 x_{2t} - \dots - \beta_n x_{nt} = e_t. \quad (5.57)$$

Here e_t must be stationary if the equilibrium is significant. Non-stationary variables can build a linear relationship using this method. The Engle-Granger methodology can be used to look for cointegration by attempting to identify the stationarity of the equilibrium relationship's residuals. To begin, the unit root test outlined in the preceding section is used to establish the order of integration of each variable. Because we are ignoring multicointegration, which is beyond the scope of this thesis, all variables should be integrated in the same order. The long-run equilibrium connection is then calculated

using the equation:

$$x_t = \alpha_0 + \alpha_1 z_t + e_t. \quad (5.58)$$

If there is cointegration, the α_0 and α_1 approximations in the OLS regression reveal "super-consistent" estimators. The fitted values of the e_t series (\hat{e}_t) are tested for stationarity in this estimation. DF or ADF tests could be utilised in this analysis. In hypothesis testing, however, critical values developed by MacKinnon (1996) are applied. We can deduce that there is cointegration between x_t and z_t if this series is stationary. \hat{e}_t can be used as the model's error correcting term.

5.15 VECM Theoretical Notions and the Model

One of the models in Multivariate Time Series is VECM (Vector Error Correction Modelling) that was proposed and analysed by Johansen (1995). ECM (Error Correction Modelling), a long-term association between particular non-stationary variables in the original data, is the most basic type of univariate modelling. The introduction of this cointegration is like a new ray of hope for the long-term establishment of a stable state using a combination of linear variables. So, if two or more non-stationary time series are integrated together in a way that they cannot deviate from equilibrium in the long term, they are considered to be cointegrated. ECM can be employed if cointegration analysis is probable. If testing reaches the ECM analysis (short-term relationship), the term Error Correction Term (ECT) will be used. This is utilised when the rate of adjustment of the state of equilibrium is projected to be negative (convergent). Furthermore, the potential of ECM in a cointegration study is equivalent to the normal regression of known terms of independent and constrained variables.

One of the specific versions of system simultaneous equations is the vector autoregression (VAR). If all of the variables are steady, VAR can be used. If the variables in vector \mathbf{Z}_t are nonstationary, the model utilised is the Vector Error Correction Model (VECM) if the variables have at least one or more cointegration relationships. The

VECM is a VAR that was created to work with nonstationary data with a cointegration connection (Enders, 2008).

In this thesis, we used the VAR and VECM models to analyse our hypothesis of how air pollution influences public health. The VECM model allows us to estimate the long and short run equilibrium effect for each variable interacting with itself and with other variables without imposing any theoretical structure on the estimates or making any assumptions regarding exogeneity of variables a priori. If all variables in our VAR co-integrate with order $I(1)$, and if there are cointegration relationships between them, then we utilise a VECM to approximate the impulse response functions. The number of co-integrating vectors is indicated by the cointegration rank in VECM. Combinations of two linearly independent, non-stationary variables with a rank of two, for example, are stationary. If the ECM coefficient is negative and substantial (e_{i-1} in the preceding equations), any short-term fluctuations between the independent variables and the dependent variable will generate a stable long-run relationship between the variables. If there is no cointegration among the variables, the result of the cointegration analysis allows us to determine whether the specific model will be a VAR model in its group form, or a VAR model in the form of a VECM, if there is at least one association of cointegration between the variables. A VAR model can be written as follows in matrix notation:

$$y_t = A_0 + A_1 y_{t-1} + \dots + A_p y_{t-p} + B_0 z_t + B_1 z_{t-1} + \dots + B_p z_{t-r} + \varepsilon_t \quad (5.59)$$

in cases where y is a $n \times 1$ vector with the model's endogenous variables, and z is a $m \times 1$ vector with the model's exogenous variables. A_0 is a $n \times 1$ vector of intercepts; A_1, \dots, A_p are $n \times n$ coefficient matrices that connect endogenous variable lag values to present values; B_1, \dots, B_p are $n \times m$ coefficient matrices that connect exogenous variable current values to endogenous variable values; and e_t is a $n \times 1$ vector of random disturbances IID $N(0, \sigma^2)$.

Because of the system's cointegration linkages, a Vector of Error Correction Model (VECM) must be utilised instead of a VAR model. Granger and Engle (1987) developed VECM models with the goal of including short-term modifications due to the presence of cointegration. The following is a representation of a VECM model:

$$\Delta y_t = \Pi_1 y_{t-k} + \Gamma_1 \Delta y_{t-1} + \Gamma_2 \Delta y_{t-2} + \dots + \Gamma_{k-1} \Delta y_{t-(k-1)} + u_t. \quad (5.60)$$

where $\Pi = \left(\sum_{j=1}^k \beta_j \right) - I_g$; $\Gamma_i = \left(\sum_{j=1}^i \beta_j \right) - I_g$, Δy_t is a vector of differences with n variables, $u_t \sim (0, \Sigma)$, Σ is a covariance matrix of u_t with $E(u_t u_s') = 0 \forall t \neq s$. On the left side of the equation, there are g variables, and on the right side, there are $k - 1$ dependent variable delays, each of which is coupled with a coefficient matrix Γ_i (Johansen and Juselius, 1990).

The Vector Error Correction Model is a restriction variant of the Vector Autoregressive Model. Because there are non-stationary yet cointegrated data forms, this additional limitation is necessary. The cointegration restriction information is then used by VECM in its requirements. For nonstationary series with cointegration relationships, VECM is often referred to as the VAR design. After the cointegration has been determined, the following test is carried out using the error correction approach. If the test variables have different degrees of integration, the test is conducted simultaneously between the long-term equations and the error correction equation after it is determined that the cointegration variable exists. Lee and Granger refer to multi cointegration as the degree of integration for cointegrated variables. If no cointegration event is observed, the test is repeated using the first difference variable. The VAR is a particular instrument with a specific function in understanding the interaction between model variables. Forecast Error Variance Decompositions (FEVD) and Impulse Response Function (IRF) often known as Variance Decompositions (VD), are two of the instruments (Lütkepohl, 2005).

5.16 VAR and VEC Models

When it's unclear whether or not a variable in an equation is exogenous, it's best to handle it symmetrically. This means assuming that there is an x_t series that is influenced by present and previous values of z_t , as well as a variable z_t that is influenced by current and previous values of x_t . The following bi-variate system can be written in this situation:

$$\begin{aligned} x_t &= a_{11} - a_{12}z_t + \sum_{i=1}^p \tau_{1i}z_{t-i} + \sum_{i=1}^p \psi_{1i}x_{t-i} + \xi_{1t}, \\ z_t &= a_{21} - a_{22}x_t + \sum_{i=1}^p \tau_{2i}z_{t-i} + \sum_{i=1}^p \psi_{2i}x_{t-i} + \xi_{2t}. \end{aligned} \quad (5.61)$$

This is a bivariate VAR model in which x_t and z_t are assumed to be stationary and have white-noise and the error factors are uncorrelated. Since x_t is correlated with z_t and ξ_{2t} is correlated with ξ_{1t} , the system of equations produced by (5.61) cannot be approximated directly. When using typical estimation approaches, regressors should not be correlated with the error term. A reduced variant of the VAR model is built in order to estimate the VAR model. We get a simplified form of the VAR model after making the appropriate corrections:

$$\begin{aligned} x_t &= \alpha_{11} + \sum_{i=1}^p \lambda_{1i}z_{t-i} + \sum_{i=1}^p \delta_{1i}x_{t-i} + e_{1t}, \\ z_t &= \alpha_{21} + \sum_{i=1}^p \lambda_{2i}z_{t-i} + \sum_{i=1}^p \delta_{2i}x_{t-i} + e_{2t}. \end{aligned} \quad (5.62)$$

The error terms e_{1t} and e_{2t} in the simplified form of the system are composites of the two shocks ξ_{1t} and ξ_{2t} . Both e_{1t} and e_{2t} have constant variances, zero means, and are serially uncorrelated since ξ_{1t} and ξ_{2t} are white-noise processes. If the lag length is 1 ($p = 1$), the error terms e_{1t} and e_{2t} are as follows:

$$\begin{aligned} e_{1t} &= (\xi_{1t} - a_{12}\xi_{2t}) / (1 - a_{12}a_{21}), \\ e_{2t} &= (\xi_{2t} - a_{21}\xi_{1t}) / (1 - a_{12}a_{21}). \end{aligned} \quad (5.63)$$

We should test if one of the lagged endogenous variables has an effect on the other endogenous variable after estimating this system of equations (5.62). The conventional F-test is used to examine this under the assumption of variable stationarity. When determining if z_t has an effect on x_t , the null hypothesis is $H_0 : \lambda_{1i} = 0$, and the alternative hypothesis is $H_a : \text{one of the } \lambda_{1i} \text{ is not zero, where } i = 1, 2, \dots, p$. Similarly, when determining if x_t has an effect on z_t , the null hypothesis is $H_0 : \delta_{1i} = 0$, and the alternative hypothesis is $H_a : \text{one of the } \delta_{1i} \text{ is not equal to zero, where } i = 1, 2, \dots, p$. We can conclude that z_t has an effect on x_t if the null hypothesis is rejected.

Because of the presence of a cointegrating association, a linear combination of non-stationary variables may be stationary, as previously mentioned. The error term must be stationary if the long-run equilibrium is relevant. Instead of a differenced VAR model, a Vector Error Correction Model (VECM) might be created with this functionality. The VECM is illustrated by the following model:

$$\begin{aligned}\Delta x_t &= \mu_{11} + \sum_{i=1}^{p-1} \theta_{1i} \Delta z_{t-i} + \sum_{i=1}^{p-1} \gamma_{1i} \Delta x_{t-i} + \beta_{11} EC_{t-1} + u_{1t}, \\ \Delta z_t &= \mu_{21} + \sum_{i=1}^{p-1} \theta_{2i} \Delta z_{t-i} + \sum_{i=1}^{p-1} \gamma_{2i} \Delta x_{t-i} + \beta_{21} EC_{t-1} + u_{2t}.\end{aligned}\tag{5.64}$$

In Eq. (5.64), Δx_{t-i} and Δz_{t-i} are stationary variables, where EC_{t-1} is the error correction term, and u_{1t} and u_{2t} are stationary error terms. The short-run dynamics of the variables in an error correction model are influenced by the departure from equilibrium. The coefficients of the lagged right hand side variables ($\theta_{1i}, \theta_{2i}, \gamma_{1i}$ and γ_{2i}) demonstrate a short run effect, which is referred to as the impact multiplier in this methodology. The adjustment effects are the coefficients of the error correction variables (β_{11} and β_{21}), which show the correction of the disequilibrium from the long-run equilibrium. When the coefficient of the error correction term is big, the response to the prior period's deviation from long-run equilibrium is considerable, whereas when the coefficient is small, the left hand side variable is unresponsive to the previous period's equilibrium error (Becker et al., 2004). The adjustment coefficient in this model captures the long-run relationship. If both error correction term coefficients are 0, we deduce that there is no

long-run link and the model should be approximated using the VAR model. If one of the adjustment coefficients in a VECM is 0, the other adjustment coefficient handles all of the adjustments. The endogenous variable with a zero adjustment coefficient might be viewed as weakly exogenous in this scenario.

Cointegration must be discovered before VECM can be used. Four phases are proposed by the Engle-Granger methodology. The unit root test presented in this Chapter is used to establish the order of integration of each variable in the first step. In the second stage, Equation (5.58) is used to estimate the long-run equilibrium association, and the fitted values of the error term e_t series are checked for stationarity by comparing the ADF or DF test statistic to critical values (MacKinnon, 1996).

We can deduce that there is cointegration between x_t and z_t if this series is stationary. Engle and Granger (1987) suggested the VECM as an instrumental variable for the $(x_{t-1} - \alpha_1 z_{t-1})$, and \hat{e}_t can be utilised as an error correction term of the VECM. Equations (5.64) form the VECM, with EC_{t-1} being \hat{e}_t . The final phase involves estimating VECM and determining the significance of each coefficient of lagged endogenous variables and coefficient of error correction terms. The F-test can be used to test the restriction on lagged endogenous variables' coefficients, and the t-test can be used to evaluate the significance of adjustment coefficients. The model's appropriateness should be tested as the final phase.

The lag duration in both VAR and VEC models can be found by utilising the conventional VAR model in levels. The VAR model is evaluated using various lag length selection methods such as the Hannan-Quinn Information Criterion (HQ), Schwarz Information Criterion (SC), Final Prediction Error (FPE), Sequential Modified Likelihood Ratio (LR) and Akaike Information Criterion (AIC). The outcomes of these five data criteria may be contradictory. The criterion is determined based on the theory and prior knowledge about the relationship in question, as the goal is to find the best feasible outcomes. In most cases, the Akaike Information Criterion (AIC) or the Schwarz Bayesian Criterion is employed to determine the length of the lag. The optimal model is chosen as the one that minimizes these criteria.

5.17 Cointegration and VECM

When two series are identified as co-integrated, it is assumed that they have a long-term equilibrium association, thus we use VECM to assess the co-integrated series' short-run characteristics. We bypass VECM and move right to Granger causality tests to discover the causal links between variables if there is no cointegration. The VECM regression equation looks like this:

$$\begin{aligned}\Delta Y_t &= \alpha_1 + p_1 e_t + \sum_{i=0}^n \beta_i \Delta Y_{t-i} + \sum_{i=0}^n \delta_i \Delta X_{t-i} + \sum_{i=0}^n \gamma_i Z_{t-i}, \\ \Delta X_t &= \alpha_2 + p_2 e_{t-1} + \sum_{i=0}^n \beta_i Y_{t-i} + \sum_{i=0}^n \delta_i \Delta X_{t-i} + \sum_{i=0}^n \gamma_i Z_{t-i}.\end{aligned}\tag{5.65}$$

The number of co-integrating vectors is represented by the cointegration rank in the VECM. A rank of two, for instance, shows that two non-stationary variable amalgamations will be stationary if they are linearly independent. If the ECM coefficient is negative and significant (i.e. e_{t-1} in the above equations), any short-term oscillations between the independent and dependent variables will result in a steady long-run relationship between the variables.

- **JOHANSEN'S METHODOLOGY FOR MODELLING COINTEGRATION** The basic steps in Johansen's methodology are:

- Specify and estimate a VAR(p) model for \mathbf{Y}_t ;
- Construct likelihood ratio tests for the rank of $\mathbf{\Pi}$ to determine the number of cointegrating vectors;
- If necessary, impose normalisation and identifying restrictions on the cointegrating vectors;
- Given the normalised cointegrating vectors estimate the resulting cointegrated VECM by maximum likelihood.

5.18 Specification of Deterministic Terms

Considering the cointegrated VECM for $\Delta \mathbf{Y}_t$, that can be written as:

$$\Delta \mathbf{Y}_t = \boldsymbol{\alpha} \boldsymbol{\beta}' \mathbf{Y}_{t-1} + \boldsymbol{\varepsilon}_t. \quad (5.66)$$

The VECM formula equation by equation gives:

$$\begin{aligned} \Delta y_{1t} &= \alpha_1 (y_{1t-1} - \beta y_{2t-1}) + \varepsilon_{1t}, \\ \Delta y_{2t} &= \alpha_2 (y_{1t-1} - \beta y_{2t-1}) + \varepsilon_{2t}. \end{aligned} \quad (5.67)$$

The change in y_{1t} according to the lagged disequilibrium error $\boldsymbol{\beta}' \mathbf{Y}_{t-1} = (y_{1t-1} - \beta y_{2t-1})$ is explained in the first equation, however, the second equation explains the change in the Δy_{2t} to the lagged disequilibrium error as well. Notice that the reactions of y_1 and y_2 to the disequilibrium errors are captured by the adjustment coefficients α_1 and α_2 .

If the deterministic terms are unrestricted, then the time series in \mathbf{Y}_t may exhibit quadratic trends and there may be a linear trend term in the cointegrating relationships. Restricted versions of the trend parameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ limit the trending nature of the series in \mathbf{Y}_t . The trend behaviour of \mathbf{Y}_t can be classified into five cases:

1. Model $H_2(r) : \boldsymbol{\mu}_t = 0$ (no constant). The restricted VECM is:

$$\Delta \mathbf{Y}_t = \boldsymbol{\alpha} \boldsymbol{\beta}' \mathbf{Y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \mathbf{Y}_{t-1} + \cdots + \boldsymbol{\Gamma}_{p-1} \Delta \mathbf{Y}_{t-p+1} + \boldsymbol{\varepsilon}_t \quad (5.68)$$

and all the series in \mathbf{Y}_t are $I(1)$ without drift and the cointegrating relations $\boldsymbol{\beta}' \mathbf{Y}_t$ have mean zero.

2. Model $H_1^*(r) : \boldsymbol{\mu}_t = \boldsymbol{\mu}_0 = \boldsymbol{\alpha} \boldsymbol{\rho}_0$ (restricted constant). The restricted VECM is

$$\Delta \mathbf{Y}_t = \boldsymbol{\alpha} (\boldsymbol{\beta}' \mathbf{Y}_{t-1} + \boldsymbol{\rho}_0) + \boldsymbol{\Gamma}_1 \Delta \mathbf{Y}_{t-1} + \cdots + \boldsymbol{\Gamma}_{p-1} \Delta \mathbf{Y}_{t-p+1} + \boldsymbol{\varepsilon}_t. \quad (5.69)$$

The series in \mathbf{Y}_t are $I(1)$ without drift and the cointegrating relations $\boldsymbol{\beta}' \mathbf{Y}_t$ have non-zero means $\boldsymbol{\rho}_0$

3. Model $H_1(r) : \boldsymbol{\mu}_t = \boldsymbol{\mu}_0$ (unrestricted constant). The restricted VECM is:

$$\Delta \mathbf{Y}_t = \boldsymbol{\mu}_0 + \boldsymbol{\alpha} \boldsymbol{\beta}' \mathbf{Y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \mathbf{Y}_{t-1} + \cdots + \boldsymbol{\Gamma}_{p-1} \Delta \mathbf{Y}_{t-p+1} + \boldsymbol{\varepsilon}_t. \quad (5.70)$$

The series in \mathbf{Y}_t are $I(1)$ with drift vector $\boldsymbol{\mu}_0$ and the cointegrating relations $\boldsymbol{\beta}' \mathbf{Y}_t$ may have a non-zero mean.

4. Model $H^*(r) : \boldsymbol{\mu}_t = \boldsymbol{\mu}_0 + \boldsymbol{\alpha} \boldsymbol{\rho}_1 t$ (restricted trend). The restricted VECM is:

$$\begin{aligned} \Delta \mathbf{Y}_t = & \boldsymbol{\mu}_0 + \boldsymbol{\alpha} (\boldsymbol{\beta}' \mathbf{Y}_{t-1} + \boldsymbol{\rho}_1 t) \\ & + \boldsymbol{\Gamma}_1 \Delta \mathbf{Y}_{t-1} + \cdots + \boldsymbol{\Gamma}_{p-1} \Delta \mathbf{Y}_{t-p+1} + \boldsymbol{\varepsilon}_t. \end{aligned} \quad (5.71)$$

The series in \mathbf{Y}_t are $I(1)$ with drift vector $\boldsymbol{\mu}_0$ and the cointegrating relations $\boldsymbol{\beta}' \mathbf{Y}_t$ have a linear trend term $\boldsymbol{\rho}_1 t$.

5. Model $H(r) : \boldsymbol{\mu}_t = \boldsymbol{\mu}_0 + \boldsymbol{\mu}_1 t$ (unrestricted constant and trend). The unrestricted VECM is:

$$\Delta \mathbf{Y}_t = \boldsymbol{\mu}_0 + \boldsymbol{\mu}_1 t + \boldsymbol{\alpha} \boldsymbol{\beta}' \mathbf{Y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \mathbf{Y}_{t-1} + \cdots + \boldsymbol{\Gamma}_{p-1} \Delta \mathbf{Y}_{t-p+1} + \boldsymbol{\varepsilon}_t. \quad (5.72)$$

The series in \mathbf{Y}_t are $I(1)$ with a linear trend (quadratic trend in levels) and the cointegrating relations $\boldsymbol{\beta}' \mathbf{Y}_t$ have a linear trend.

5.19 Impulse Response Functions and Variance Decompositions

Innovation accounting, which consists of impulse response and variance decomposition analysis, is used to represent system dynamics. In VAR, variance decomposition decomposes variation in an endogenous variable into component shocks to the endogenous variable, whereas an impulse response function traces the effect of one standard deviation shock to one of the innovations on current and future values of the endogenous variables.

When considering equations (5.62), a shock to a variable affects the variable itself. Because VAR has a dynamic structure, this effect is passed on to all of the system's endogenous variables. Because of the presence of lagged x_t in both equations, a change in e_{t1} will have an immediate effect on x_t and will also impact future values of z_t and x_t . If the innovations e_{t1} and e_{t2} are not correlated, e_{t1} represents x_t and e_{t2} represents z_t . However, in practice, the innovations are frequently connected, resulting in a shared component between the two variables that cannot be associated with one of them individually. This issue could be overcome by attributing the entire effect of any common component to the VAR system's initial variable. Cholesky decomposition is the term given to this procedure. This analysis can vary depending on the order of the variables in the VAR system; as a result, this property should be considered while performing an impulse response analysis. To put it another way, variance decomposition "provides the variance of a given variable's forecast errors to its own shocks and the shocks of the other variables in the VAR model" (Lanne and Nyberg, 2016).

Chapter 6

The Long/Short Run Relationship Between RA and AQI Using VECM

6.1 Introduction

This chapter will present and apply the empirical idea of cointegration, however it is necessary to first introduce the concept's essence in order to comprehend its implications. The primary principle behind cointegration is that the variables have a propensity to move collectively in the long run, implying that they have reached a state of equilibrium. Deviations in the short term from equilibrium are conceivable, but due to the error or equilibrium correction model, the variables will recover to their equilibrium relation in the long run (Engle and Granger, 1987).

The widely utilised cointegration approach in econometrics has been found to provide approximations that give more accurate results for air pollution trend identification and attribution (Tang et al., 2019; Taghizadeh-Hesary and Taghizadeh-Hesary, 2020; Zhu et al., 2019; Zou et al., 2016). The maximum likelihood (ML) estimates from a cointegrating vector autoregressive (VAR) model are compared to the total least squares (TLS) and the conventional ordinary least squares (OLS) estimates from a static regression model.

Environmental time series can be seen in the form of air quality measurements. The most common methodology for estimating environmental parameters is conventional descriptive statistics, but due to the considerable variability associated with the low signal-to-noise ratio of the available observations and air quality data, this is of little value. Time series analysis, which allows the detection of underlying deterministic behaviour and so contributes to the understanding of cause and effect links in environmental problems, may be a viable way to avoid these difficulties (Schwartz and Marcus, 1990).

The time series forecasting method is beneficial for projecting future air quality conditions based on numerous characteristics of each country's development. To create a forecasting model, the forecasting approach examines the sequence of historical data inside a time series. The ARIMA approach has been thoroughly investigated and applied in prior research, and it has been shown to be effective in the field of forecasting. Many prior articles have discussed forecasting approaches for the pollution field using the ARIMA time series method (Abhilash et al., 2018; Wang and Guo, 2009; Kumar and Jain, 2010).

So, this chapter will present the univariate and multivariate time series analysis for predicting disease activity scores (DAS28) among patients from Kuwait with rheumatology arthritis (RA) using the information of air pollution. In addition, we will examine the long and short run relationship between air pollution components such as NO_2 , SO_2 and O_3 with the rheumatoid arthritis (RA) disease activity score (DAS28).

6.2 Background of the Study

According to a 2016 WHO report, 7 million people die each year as a consequence of being exposed to ambient (outdoor + indoor) air pollution around the world (Organization, 2016). People who are very young, elderly, have pre-existing respiratory disorders, or have a low socioeconomic status are the most vulnerable to air pollution. The pollutants with the most consequential evidence of health effects were found to be CO , PM_{10} , NO_2 ,

SO_2 and O_3 . Several studies have linked ambient air pollution to a variety of negative health effects, ranging from asymptomatic to fatal results (Dominici et al., 2006; Li et al., 2018; Mölter et al., 2015).

The relationship between RA risk and the exposure to several specific air pollutants, including NO_2 , SO_2 , and PM_{10} , from local sources (traffic and home heating) was investigated in a recent analysis from the Swedish Epidemiological Investigation of Rheumatoid Arthritis (EIRA) case-control study, and the outcomes indicate that selected pollutants (NO_2 and SO_2) had a relationship with increased risk of RA (Hart et al., 2013b). This also agreed with Chang et al. (2016) who discovered that $PM_{2.5}$ and NO_2 are risk factors of increasing RA in participants.

Also, a study in Korea suggested an increased risk of incident RA in adults exposed to CO and O_3 (Shin et al., 2019). They found a positive relationship between exposure to some gaseous pollutants (CO and O_3) and the risk of RA.

A study was undertaken in Kuwait to determine the relationship between disease activity ratings and air pollution in RA patients. The study concluded that NO_2 and SO_2 were discovered to be important risk factors for the development of RA (Alsaber et al., 2020).

Next, after discovering the relationship between the pollutions (NO_2 , SO_2 and O_3) with the rheumatology disease activity, it is important to build a multivariate time series model that predicts the disease activity of RA given the information of the most correlated components of air pollution with RA (NO_2 , SO_2 and O_3).

6.3 Multivariate Time Series and Air Pollution

Various extended models such as VARMA (Vector Autoregressive Moving Average) and VAR (Vector Autoregressive) have been used for multivariate time series analysis in environmental studies. VAR and VEC (Vector Error-correction) were exploited for long-term prediction based on multivariate time series data (Cox et al., 1981), and VARMA was used for multivariate time series data in financial services (Veenstra and Haralambides,

2001). The study in Chevillon and Hendry (2005) researched prediction performance of the VAR model using direct multi-step estimation for both stationary and non-stationary time series generated in economic activities, and the study in Haldrup et al. (2010) forecasted the price of electricity by a VAR model using fractional cointegration. One drawback of the VAR model is that the number of parameters to be estimated can become large (Gong et al., 2019). Recently, predictions for fuzzy time series were performed using a multivariate heuristic model (Huarng et al., 2007), and a new method using a fuzzy relation based on a neural network algorithm was suggested for high-dimensional time series data (Egrioglu et al., 2009). However, these models need to satisfy too many conditions and constraints; the VARMA model is slow with complicated data (Isufi et al., 2019). In addition, even if the models handle multivariate time series data, they are usually not suitable to forecast a certain dependent variable of the data. Thus, given many situations, traditional single models for forecasting multivariate time series have limits.

With the fast application and development of sensor technology and the Internet of Things in the big data era, air quality prediction is becoming increasingly reliant on a variety of data acquisition equipment and sensors to gather big data for urban air, such as weather data, PM_{10} , NO_2 , $PM_{2.5}$, and traffic data, among other things. Some academics have been working on air quality predictions and air pollution incidences in recent years (Zhang et al., 2012a,b). However, to characterise the evolution of air pollution, the majority of these research studies rely on mathematical equations or simulation methodologies (Vardoulakis et al., 2003). Classic shallow machine learning algorithms epitomise these conventional methodologies. In another example, Dong et al. (2009) introduced a unique approach for $PM_{2.5}$ concentration value prediction based on hidden semi-Markov models (HSMs). Based on the Nonparametric Regression approach and Integrated Parametric approach, Donnelly et al. (2015) proposed a model for creating real-time air quality forecasts with excellent accuracy and computing efficiency. Because air pollution is influenced by weather, traffic, and other factors, statistical approaches and shallow machine learning models struggle to effectively capture and predict it.

Air quality prediction has a long history of research in the literature; most of the existing studies use shallow machine learning models and statistical methods to handle the difficulties of air quality prediction (Vardoulakis et al., 2003) including Regression (Donnelly et al., 2015), ARIMA (Díaz-Robles et al., 2008), Artificial Neural Networks (Zhou et al., 2014) and HMM (Dong et al., 2009). For example, Zhang (2007) offered a comprehensive overview of real-time air quality forecasting difficulties, including their history, important research, current state and future directions (Zhou et al., 2015; Zheng et al., 2013). To identify the dynamic temporal relationships of $PM_{2.5}$, Zhang (2007) developed a probabilistic dynamic causal (PDC) model based on Lasso-Granger to uncover the dynamic temporal dependencies (Zhou et al., 2015). Founded on ensemble empirical mode decomposition and a general regression neural network technique, Zhang (2007) built a hybrid model for one-day-ahead $PM_{2.5}$ prediction (Zhou et al., 2014). In another example, Deleawe et al. studied the application of machine learning technologies to predict CO_2 levels in urban air settings, which is an indicator of air quality (Deleawe et al., 2010).

The Box and Jenkins technique is used to develop functional correlations between several time series variables and it was examined for environmental studies. Moving average, stochastic processes, autoregressive, and autoregressive integrated moving average models are commonly utilised. To cope with the non-stationarity of the time series, the Box and Jenkins technique offers differencing of the variables. However, differencing the series is not always wanted because it confines the model to only short-run fluctuations, obviating the importance of long-run variations, which may be a key aspect of the stochastic process. Furthermore, if the series act in an equilibrium manner, estimating the magnitude of the departures from the equilibrium route may be informative. Furthermore, if non-stationarity is neglected, a relationship may be constructed when none existed; erroneous conclusions can be generated from non-stationary series regression (Granger and Newbold, 1974).

As a result, the model may need to consider not just long-run but also short-run fluctuations. Though analysis can be founded solely on short-run changes, this may

not be particularly useful because long-run variations, assuming they are important, are neglected when determining a relationship. De-trended variables might also be employed to avoid non-stationarity, but the dynamic model would still only describe the long-run phenomenon. Furthermore, choosing certain variables as exogenous and others as dependent at random is not a recommended method for evaluating connected time series. As a result, a new methodology is required.

Granger proposes the cointegration strategy to deal with these kinds of problems (Engle and Granger, 1987). There are two types of cointegration models to cope with non-stationarity: the vector error correction model (VECM) and vector autoregressive (VAR). If the coefficient matrix related with the stochastic equations indicating variable associations is less than full rank, either (1) cointegration restrictions on the coefficients are applied, (2) the error correction representation is used, or (3) the random walk element is disjointed, leaving the approximation to progress in the VAR representation. Level as well as differenced variates are included in the error correction model. As a result, adopting this version of the vector autoregressive model is advantageous because it includes both short and long run parameters. These may stray from each other from period to another period, but they normally go in the same direction. They may diverge in the short term, but in the long run, they will converge. Assuming an arbitrary sequence of variables and using univariate models may not be acceptable for such time series because they could be endogenous at the same time. Instead, at least as a starting point for the inquiry, a systems or simultaneous equations method could be more applicable. Surprisingly, a cointegration strategy using a mixture of stochastic equations produces a stable process even in situations where the variables are non-stationary.

According to what we have presented, in the next section we will present the importance of this study in term of implementing a multivariate time series approach in order to capture and predict the influence of air pollutants toward rheumatology disease activity scores among RA patients from Kuwait.

6.4 Importance of the Study

This study is limited to cover the situation in Kuwait as a country. Over the previous four decades, Kuwait's infrastructure and socioeconomic development has been extremely rapid. Hundreds of kilometres of metropolitan highways and arterials have been built to accompany the rapid growth of the socioeconomic sectors. Kuwait City, with a population of over four and half million people and a vehicle fleet of over two million, is facing increased traffic volumes, increased trip frequency, and increased journey length (Al-Mutairi et al., 2009). As a result of its use for internal ventilation, outdoor air quality is becoming a serious air pollution issue and concern for residents of Kuwait City. It is also one of the most pressing issues in the urban environment. The quantities of non-methane hydrocarbons in the ambient air are generally higher than Kuwait Environment Public Authority (K-EPA) guidelines (Al-Awadhi and Al-Awadhi, 2006); suggesting oil related factors are the contributing agents. Nearly 29% of the cities tested had sulphur dioxide concentrations (typically from power plants) that were above WHO maximums, while 71% had nitrogen dioxide concentrations (primarily from urban traffic) that surpassed WHO limits (Al-Mutairi et al., 2009).

So, in this study, we want to measure the long and short term effect from air pollutants (NO_2 , SO_2 , O_3) toward chronic diseases. The rheumatoid arthritics (RA) among patients from Kuwait was chosen to be a case study for this research as a chronic disease factor. However, air pollution information was collected through four different monitoring fixed stations (see figure 6.1). In this chapter, a comparison of traditional shallow time series models and Vector Error Corrected Model (VECM) is used. By performing feature selection automatically and leveraging multivariate time series data, this is inspired to handle long temporal dependency concerns and local trend features. We presume that air quality and chronic diseases move in lockstep, so they can't deviate from one other on their own. Alternatively, any imbalance between them is a deviation in the short run (Arize, 2017).

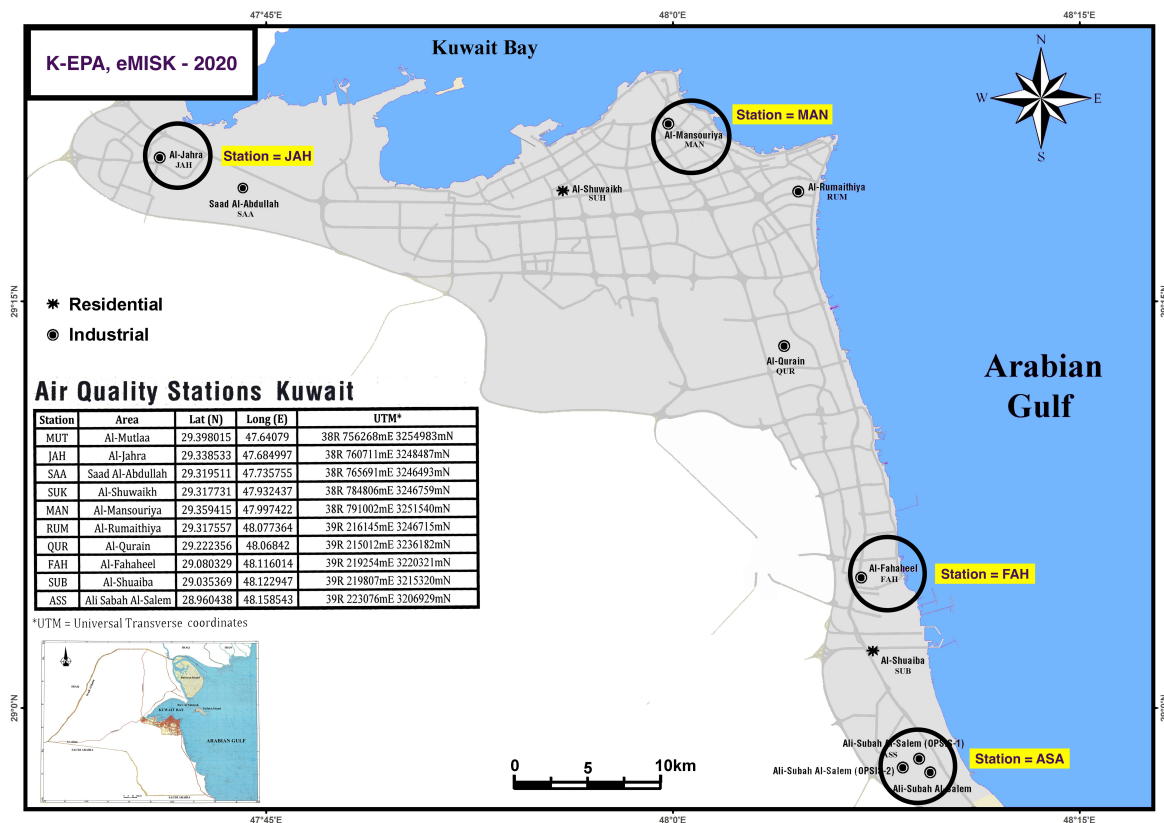


Figure 6.1: The location of monitoring fixed stations selected for this study (ASA, FAH, JAH and MAN) that belong to Kuwait Environment Public Authority (K-EPA)

6.5 Aim and Objective

As mentioned in previous chapters, the importance of the estimated coefficient of the long-run associations between model projections of air pollution variables and observations is used to detect and attribute changes in the disease scores among RA patients from Kuwait.

However, estimating and testing the predicted long-run relationship should not be considered a goal in and of itself. The two series (modelled and observational) should have a true long-run connection. Examining the function of air pollution components in the assessment of disease activity score of the patients with high Rheumatoid Arthritis

Disease Activity Score is an important subject to address before estimation (e.g. DAS28 calculated score).

6.6 Procedures and Methodology

6.6.1 Selected Variables

In this section, we will express the time series plots for the selected air pollution component in this study. As we mentioned before, the selected air pollution variables are NO_2 , SO_2 , and O_3 in addition to patients with RA disease activities score DAS28 based on their hospital records.

6.6.2 Rheumatoid Arthritis (RA) Patients' Data

Rheumatoid arthritis (RA) refers to an inflammatory illness that mostly affects the joints, producing inflammation, discomfort, and difficulty in moving them. Although the specific etiology is uncertain, the condition has been related to a number of hereditary and environmental variables. Chemical exposure was in the past suggested as a possible, if not primary, cause of the condition (Chang et al., 2016).

RA Data from KRRD

All RA patients in this study were officially registered with the Kuwait Registry for Rheumatic Diseases from January 1, 2013 to December 30, 2020. (KRRD). The KRRD is a national registry for patients with adult rheumatic illness. Patients with RA who satisfied the American College of Rheumatology (ACR) criteria (Aletaha et al., 2010) and were registered between January 2013 and December 2020 were included in the study.

Based on patient visits, RA data was obtained from the rheumatology sections of four main Kuwaiti government hospitals. The chosen hospitals are primarily located in several governorates to accommodate Kuwait's ethnic mixture. The Ethics Committees at Kuwait University's Faculty of Medicine and the Ministry of Health both approved

the KRRD, from which this study arose. In addition, all involved patients enrolled in the registry gave their official consent (Al-Herz et al., 2016). The total number of patients included in the study was 1,809 RA patients having 10,215 follow-up visits.

Calculating RA Indices

The DAS28 and CDAI indices (RA golden standards) are used to assess the severity of RA disease activity (Salaffi et al., 2009; Aletaha and Smolen, 2005; Muñoz et al., 2017). These involve the following: TJC28: The number of tender joints (0 – 28); SJC28: The number of swollen joints (0 – 28); D ESR: erythrocyte sedimentation rate (in mm/h); CRP: C-reactive protein (CRP) may be used instead of ESR in the calculation; and GH: Global health assessment of the patient (from 0 = best to 100 = worst) (see Equation (6.1)).

$$DAS28 = 0.56 \times \sqrt{TJC28} + 0.28 \times \sqrt{SJC28} + 0.70 \times \ln(ESR \text{ Or } CRP) + 0.014 \times GH. \quad (6.1)$$

Figure 6.2 shows the assessment form to assess an RA patient in order to calculate the disease activity score for the RA patient (DAS28):

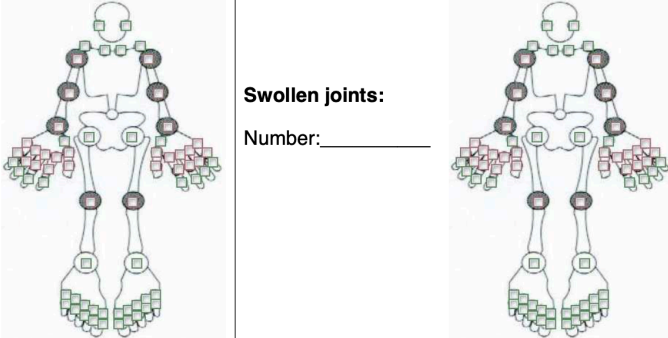
Tender joints: Number: _____	Swollen joints: Number: _____			
				
Deformities: <input type="checkbox"/> Yes <input type="checkbox"/> No Other Physical findings: _____				
Blood Test Results:				
ESR	CRP	WBC	Hgb	PLT
Creatinine	FBS	AST	ALT	ALP
T.Chol	LDL	HDL	TG	Uric acid
Urine routine: <input type="checkbox"/> Normal <input type="checkbox"/> Abnormal _____		T-Sharp score: _____	HAQ: _____	
Comments: _____		Comments: _____		
Patient's global assessment of disease activity: _____ (0=Inactive, 10=Very Active)				
Physician's global assessment of disease activity: _____ (0=Inactive, 10=Very Active)				
RADAI: _____ Is this patient pregnant? <input type="checkbox"/> Yes <input type="checkbox"/> No				

Figure 6.2: RA patient visit form; the information in this form will be used for calculating DAS28 by using the formula 6.1 for the RA patient.

6.6.3 EPA Data and Materials

The investigation used data from four Kuwait Environmental Public Authority (K-EPA) Air Monitoring Stations through the Environmental Monitoring Information System of Kuwait (eMISK) which are working under the Environmental Data Management Department. The provided data was used during the time from January 2013 till December 2020 as daily base information after we aggregate all the pollutants from an hourly base to daily. In this study, out of 18 monitoring air quality fixed stations, four were chosen, which are ASA, FAH, MAN and JAH. One of the four stations was chosen to represent Mansoria (MAN), a key urban centre with considerable traffic density; an additional two districts in the neighbourhood of oil refineries, which are Fahaheel (FAH) and Ali

Abullah Al-Salem (ASA), and the fourth is an urban district named Al-Jahra (JAH) which is located in the north of Kuwait that has the highest population density among the other three urban districts. Each of the four air pollution observation stations is around 25 kilometres apart. Figure 6.3 shows the K-EPA data processing structure per monitoring stations:

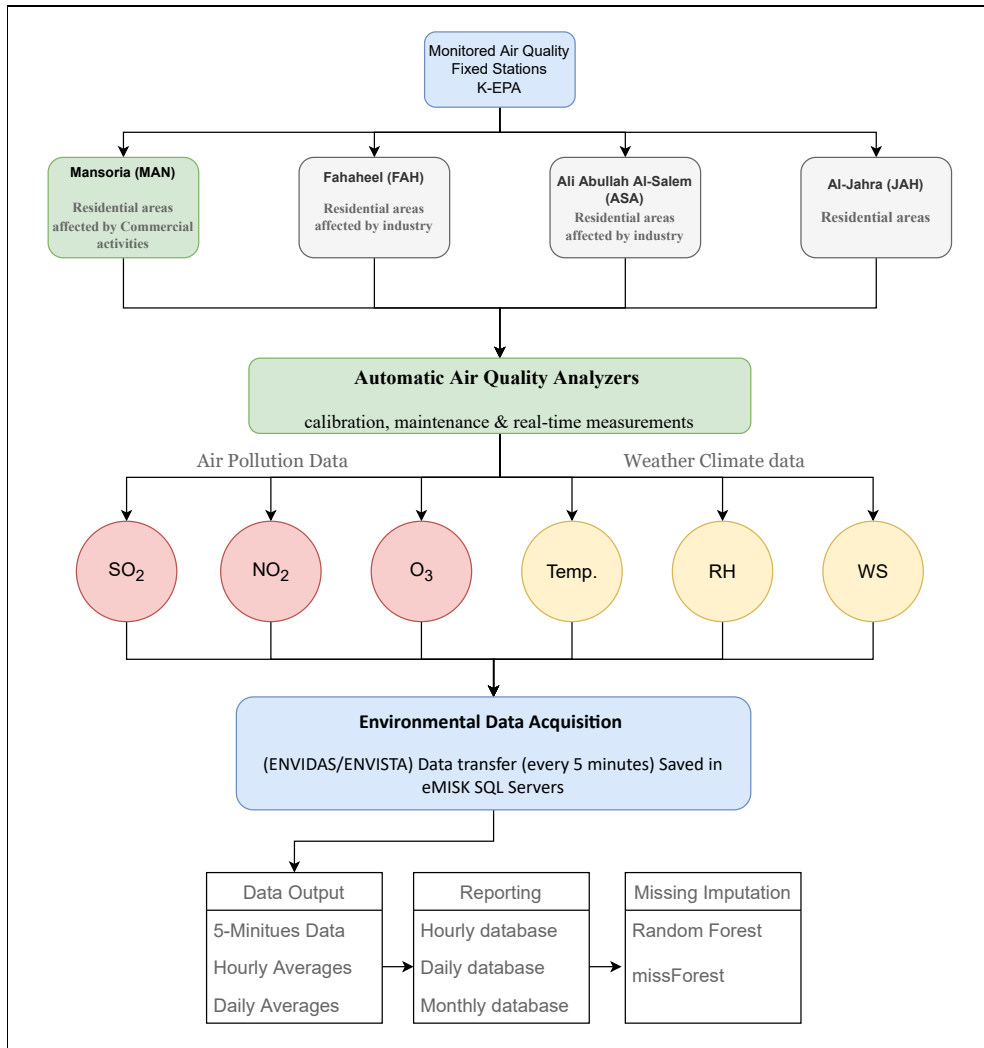


Figure 6.3: K-EPA air quality data process and flow chart.

6.6.4 Ambient Air Quality in Kuwait

The country Kuwait is situated in the north eastern corner of the Arabian Peninsula and at the tip of the Arabian Gulf. It is a small developing country covering an area of 17,818 km^2 and depends mainly on the oil and petroleum industry. Additionally, as a desert area with a scarcity of fresh water, its main source of fresh water is desalinated sea water. Kuwait hosts three main desalination plants. Furthermore, the area is affected by severe dust storms during the summer season, which highly contribute to pollution in this area (Al-Ali et al., 2020). The K-EPA maintains 18 distributed air quality monitoring stations to achieve an adequate area coverage. As we have mentioned before, four stations were selected in this study, which are ASA, FAH, MAN and JAH (Figure 6.1). The selection of these four stations was based on the observed variety of land use changes and developments, i.e., industrial and residential. This selection included the probable effect of industrial and transportation (traffic) effluents on the quality of air.

The AQI produced by Al-Shayji et al. (2008) for the State of Kuwait was used to assess air quality in this study, which was based on criteria proposed by the United States Environmental Protection Agency (USEPA) (Fitz-Simons, 1999). The air quality index (AQI) is a measurement of day-to-day air quality. It provides information about the purity of the surrounding air. As mentioned in section 1.3.5 on page 11, the AQI calculation was explained and performed using equation 1.1 on page 13 and table 1.2 on page 13.

An Air Quality Index (AQI) is a colour-coded numerical scale that is extensively used to link data on air pollution exposure to the probability of short-term unfavourable health effects. The concentrations of the common air pollutants - nitrogen dioxide (NO_2), sulphur dioxide (SO_2) and ozone - are a subset of air pollutant exposures in practice (O_3 is used to determine the AQI). R code was developed to calculate the AQI for all pollutants in the study.

For the seven-year study period, data on meteorological factors such as wind speed, temperature, humidity, and wind directions were also gathered. The data was then compiled in the cloud computing memory and processed and analysed using the following

Statistical Analysis Softwares:

RStudio - version 1.4.1717 Used for data cleaning, data manipulation, data wrangling, processing, descriptive analysis, correlation analysis, regression analysis, testing stationary level, testing for normality, multivariate time series modelling using VECM, VAR, ARIMA, GARCH and calculating the model accuracy performance using R-square, MAPE, RMSE.

Stata/SE 16.1 Used for calculating some specific results such as IRF and FEVD.

Gretl 2021a Was used to calculate the details of VECM, cointegration analysis and Granger Test.

JASP version 0.14.1 Was used to generate the formatted tables for OLS (linear regression) and Correlation analysis.

6.7 Descriptive Analysis and Correlation

Table 6.1 demonstrates the summary statistics of the study variables— Sulphur dioxide (SO_2), Nitrogen dioxide (NO_2), Ozone (O_3), temperature, relative humidity, and wind speed— for each location included in the study. The mean values of the pollutants SO_2 , NO_2 and O_3 ranged between $(5.66 \pm 2.74 - 12.67 \pm 8.85)$, $(17.18 \pm 9.27 - 29.18 \pm 13.38)$ and $(20.65 \pm 9.25 - 22.96 \pm 9.74)$ respectively. The maximum mean value of the pollutants SO_2 (12.67 ± 8.85) and NO_2 (29.18 ± 13.38) were observed in FAH station and the maximum mean value of the pollutant O_3 was observed in the MAN location with mean \pm s.d. = 22.96 ± 9.74 . The mean value of the weather parameters temperature, relative humidity and wind speed ranged between $(27.00 \pm 9.13 - 31.67 \pm 9.73)$, $(31.34 \pm 14.42 - 41.62 \pm 31.63)$ and $(2.15 \pm 0.84 - 2.55 \pm 1.33)$ respectively. The maximum mean value of the weather parameters temperature, relative humidity and wind speed was observed for the FAH (31.67 ± 9.73), JAH (41.62 ± 31.63) and ASA (2.55 ± 1.33) stations respectively.

Correlation analysis was conducted to analyse the association among the study variables, i.e. DAS28, pollutants and weather parameters (DAS28, SO_2 , NO_2 , O_3 , TEMP, RH and WS). The analysis shows significant association of DAS28 with NO_2 ($r_p = 0.029$), O_3 ($r_p = 0.039$) and WS ($r_p = 0.056$). Besides this a significant interdependence was also observed among the weather parameters and pollutants (Figure 6.4). The strongest significant positive correlation among these parameters is observed between SO_2 and O_3 ($r_p = 0.476$) whereas the strongest significant negative correlation among these parameters is observed between RH and Temp ($r_p = -0.517$). The result of correlation analysis is shown in Table 6.2.

Table 6.1: Descriptive statistics for study air pollutants per location in terms of AQI calculation.

Variable	ASA	FAH	JAH	MAN	All
<i>SO₂</i>					
min	0.337	0.143	0.190	1.874	0.143
25th quartile	3.070	6.450	6.680	3.790	4.110
median	4.640	10.890	10.000	5.310	6.450
75th quartile	7.360	16.610	15.500	7.100	11.000
max	127.300	76.400	76.000	45.600	127.000
mean (sd)	6.64 ± 7.29	12.67 ± 8.85	11.97 ± 7.82	5.66 ± 2.74	8.94 ± 7.79
<i>NO₂</i>					
min	2.854	0.943	0.750	2.276	0.75
25th	10.8	19.1	11.9	14.6	14
median	14.9	27.1	21.1	20.6	21.5
75th	21.5	36.9	32.9	29.2	32
max	95.5	99.7	153.0	111.3	153
mean (sd)	17.18 ± 9.27	29.18 ± 13.38	25.35 ± 19.60	23.65 ± 13.27	25.23 ± 16.34
<i>O₃</i>					
min	0.926	0.792	1.134	1.271	0.792
25th	16.2	14.0	14.8	16.8	15.6
median	21.5	19.3	20.3	21.8	20.6
75th	28.2	26.1	26.8	27.1	26.9
max	75.5	52.0	84.3	84.3	116
mean (sd)	22.74 ± 9.30	20.65 ± 9.25	21.36 ± 9.64	22.96 ± 9.74	22.21 ± 10.41
Temperature					
min	0.00	6.85	0.00	0.00	0
25th	17.4	23.6	18.5	17.6	19.4
median	27.8	32.6	27.6	28.3	28.6
75th	37.0	39.4	35.3	36.7	37.2
max	50.4	53.7	44.6	47.1	53.7
mean (sd)	27.08 ± 11.21	31.67 ± 9.73	27.00 ± 9.13	27.07 ± 10.19	28.17 ± 10.19
RH					
min	0.00	6.89	0.00	0.00	0
25th	18.56	17.21	3.56	23.96	21.9
median	38.1	29.3	40.3	30.5	33.8
75th	57.8	47.7	68.0	35.1	52.4
max	149.1	96.7	99.2	102.8	149
mean (sd)	40.81 ± 24.97	33.94 ± 18.83	41.62 ± 31.63	31.34 ± 14.42	38.12 ± 22.76
WS					
min	0.000	0.537	0.000	0.000	0
25th	1.58	1.54	1.74	1.12	1.46
median	2.22	1.99	2.19	1.90	2.02
75th	3.16	2.59	2.82	3.12	2.8
max	9.43	5.63	5.55	11.45	11.4
mean (sd)	2.55 ± 1.33	2.15 ± 0.84	2.34 ± 0.81	2.40 ± 1.75	2.28 ± 1.21

Table 6.2: The correlation test between DAS28, SO_2 , NO_2 , O_3 , Temp, RH and WS using Pearson's Correlations

Variable	DAS28	SO_2	NO_2	O_3	Temp	RH	WS
1. DAS28	–						
2. SO_2	0.012	–					
3. NO_2	0.029*	0.368***	–				
4. O_3	0.039**	0.476***	0.223***	–			
5. Temp	0.022	-0.020*	0.011	0.181***	–		
6. RH	0.013	-0.119***	0.016	-0.196***	-0.517***	–	
7. WS	0.056***	0.022*	-0.086***	0.236***	0.143***	-0.109***	–

* $p < .05$, ** $p < .01$, *** $p < .001$

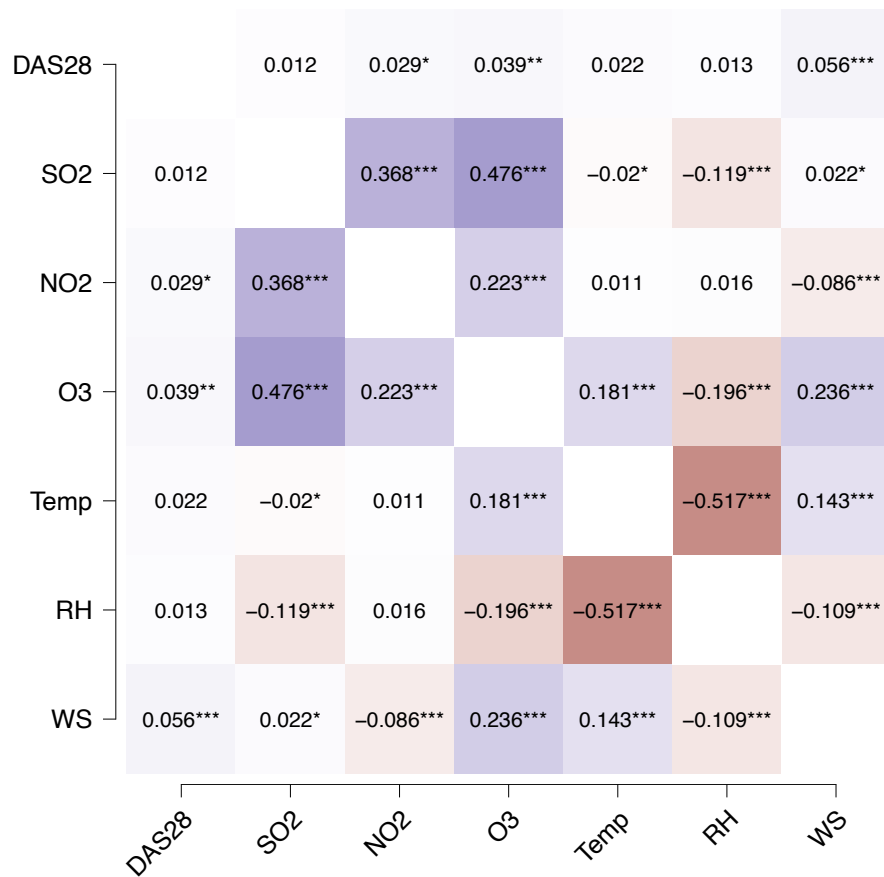


Figure 6.4: A Pearson correlation coefficient heat map between RA disease activity scores (DAS28) and the daily average concentrations of SO_2 , NO_2 , O_3 , TEMP, RH and WS. Note that * $p < .05$, ** $p < .01$, *** $p < .001$

6.8 Normality and Transformation Approach

In regression analysis, transformations are crucial (Cook and Weisberg, 1999). A transformation from a parametric family of transformations is frequently chosen. The most common power family is the Box-Cox power family (Sakia, 1992; Weisberg, 2001; Hos-sain, 2011), which is described by

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1) / \lambda & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases} \quad (6.2)$$

where y is a list of n strictly positive values. The Box-Cox family is valuable because it is identical to the power transformation family, which makes the parameter λ is easy to understand, and it contains the key special instances of untransformed logarithmic, cube root, inverse and square. The Box-Cox family is utilised in a variety of locations, including transforming a set of predictors and selecting response transformations toward multivariate normality. There have been several attempts to create transformation family variables with negative values. Consider transformations of the type $(y + \gamma)^\lambda$, where γ is large enough to ensure that the result is strictly positive. In theory, (γ, λ) might be estimated at the same time, but in fact, estimates of γ are highly varied. Other transformation families, such as the folded power family (see Cook and Weisberg, 1999, p. 330), have also been proposed, although they are rarely employed due to the poor qualities of the resulting transformations. Yeo and Johnson (2000) presented a new family of distributions that have many of the positive qualities of the Box-Cox power family and can be utilised without limits. The following are the characteristics of these transformations:

$$y_i^{(\lambda)} = \begin{cases} ((y_i + 1)^\lambda - 1) / \lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y_i + 1) & \text{if } \lambda = 0, y \geq 0 \\ - [(-y_i + 1)^{(2-\lambda)} - 1] / (2 - \lambda) & \text{if } \lambda \neq 2, y < 0 \\ - \log(-y_i + 1) & \text{if } \lambda = 2, y < 0 \end{cases} \quad (6.3)$$

The Yeo-Johnson transformation is identical to the Box-Cox power transformation of $(y+1)$ if y is strictly positive. If y is strictly negative, the Yeo-Johnson transformation is the same as the Box-Cox power transformation of $(-y+1)$, but with power $2-\lambda$. Because the transformation is a combination of both negative and positive values, different powers are needed for positive and negative values. The transformation parameter in this situation is difficult to grasp because it has different meanings for $y \geq 0$ and $y < 0$. Although the Yeo-Johnson transformation parameter is difficult to read, this family of transformations can be beneficial as techniques for choosing a transformation for linearity or normalcy (Yeo and Johnson, 2000).

6.8.1 Cullen and Frey Graph

Before developing models, different transformations could be used to improve the distribution and minimize the variability of the data. We employ descriptive statistics to choose candidate theoretic distributions or models to fit the trajectory data in this study. Kurtosis and skewness are two commonly used coefficients in descriptive statistical analysis. Kurtosis is a measure of a distribution's tailedness in comparison to the normal distribution, which has a kurtosis of 3. The term "leptokurtic" refers to distributions with a kurtosis larger than 3, whereas "platykurtic" refers to distributions with a kurtosis less than 3. The degree of asymmetry of a distribution on its mean is measured by skewness. When a distribution's skewness is positive, the probability density function on the right of the mean is "fatter" than the one on the left. The negative skewness of a distribution, on the other hand, indicates that the left part of the density function is "larger" than the right.

The following results show the pollutants' distribution and the suggested transformation approach using Cullen and Frey graphs (Figure 6.5):

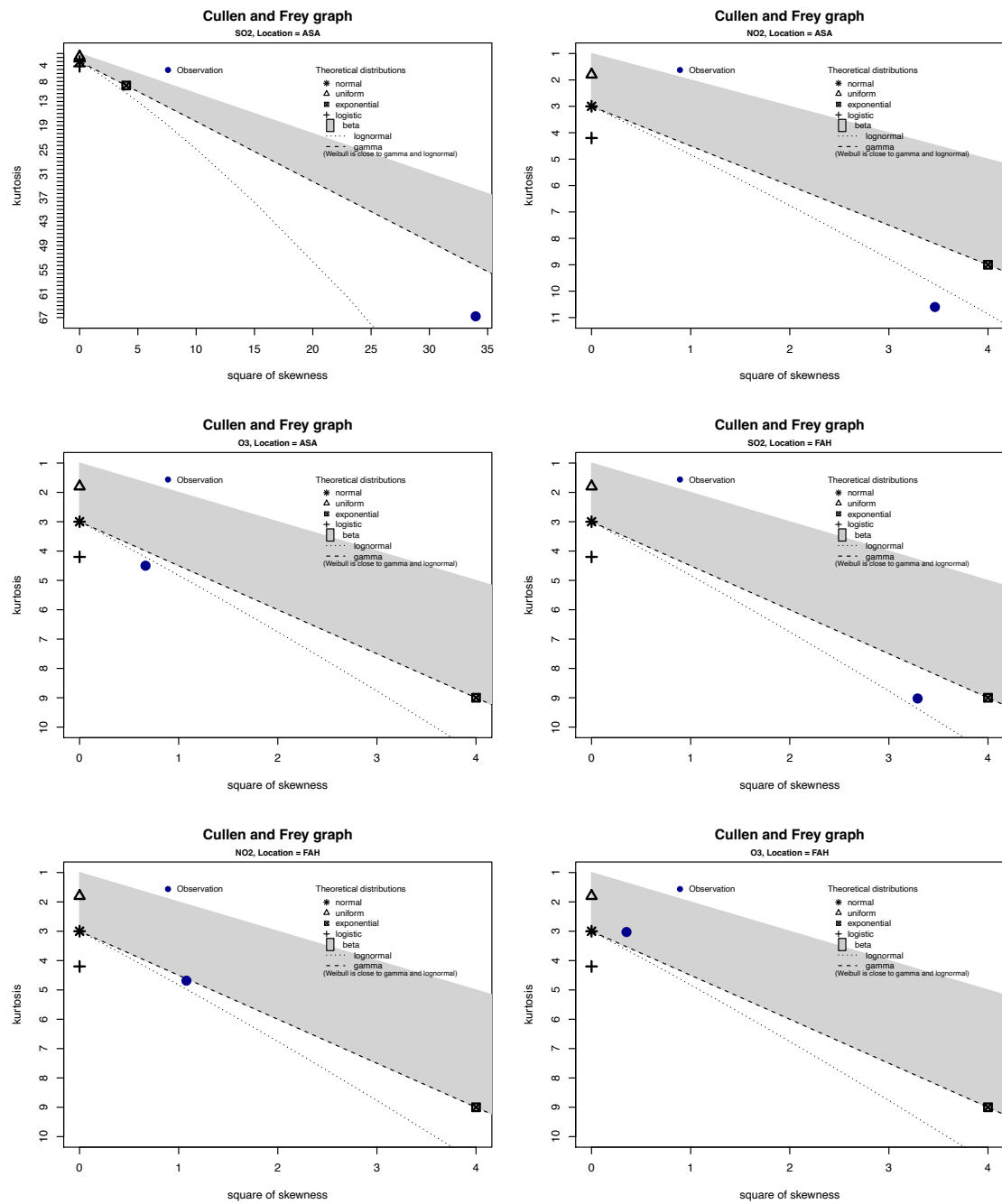


Figure 6.5: The distribution performance for Air quality index (AQI) using Cullen and Frey graphs for five governorate monitoring fixed stations in Kuwait during the period from 2013 to 2020. This figure captured the distribution performance for SO_2 , NO_2 and O_3 monitored by ASA and FAH fixed stations.

Figure 6.5, for ASA and JAH locations, shows the square of skewness of displacements and the blue circle observation is the kurtosis of the variable. For example, the first sub-figure in the left corner is related to the skewness analysis for SO_2 in the ASA location. The results indicate that the distribution of SO_2 violates the normality assumption, and the values of SO_2 require to be transformed into another shape of data. All the results for the other stations are in table 6.3.

In this study, we have implemented many transformation methods in order to reach for the best normality shape of SO_2 . Table 6.3 presents the skewness results for each pollutant in each location. We can see from table 6.3 that SO_2 in ASA has a better normality shape when we transform the values using Box-Cox transformation because the skewness result for this approach equals 6.511, that is the lowest value among the other transformation methods. However, the best transformation method for SO_2 in JAH is Yeo-Johnson with skewness results equal to 1.951, which is the lowest among the other transformation methods for SO_2 in JAH location. But, the best transformation method for SO_2 in MAN is log transform with skewness results of 42.530, which is the lowest among the other transformation methods for SO_2 in the MAN location.

We have chosen the best transformation approach for each pollutant in each location in order to promote the data for better normality performance. Figure 6.8 shows the skewness results after implementing transformation, and the distribution performance for the air quality index (AQI) using Cullen and Frey graphs in each monitoring fixed station during the period from 2013 to 2020. As we can see from figure 6.8, the solid blue circle observation has now shifted to the best normality standing point (i.e. if the solid blue circle observation becomes closer to the star point in the figure, this means that the skewness becomes closer to one).

So, figures 6.6 to 6.10 show the normality performance before and after the transformation approach in the ASA, FAH, JAH and MAN locations for SO_2 , NO_2 and O_3 . It is very obvious that the selected transformation improves the normality performance for the study variables. The results for the other locations are presented in appendix C.

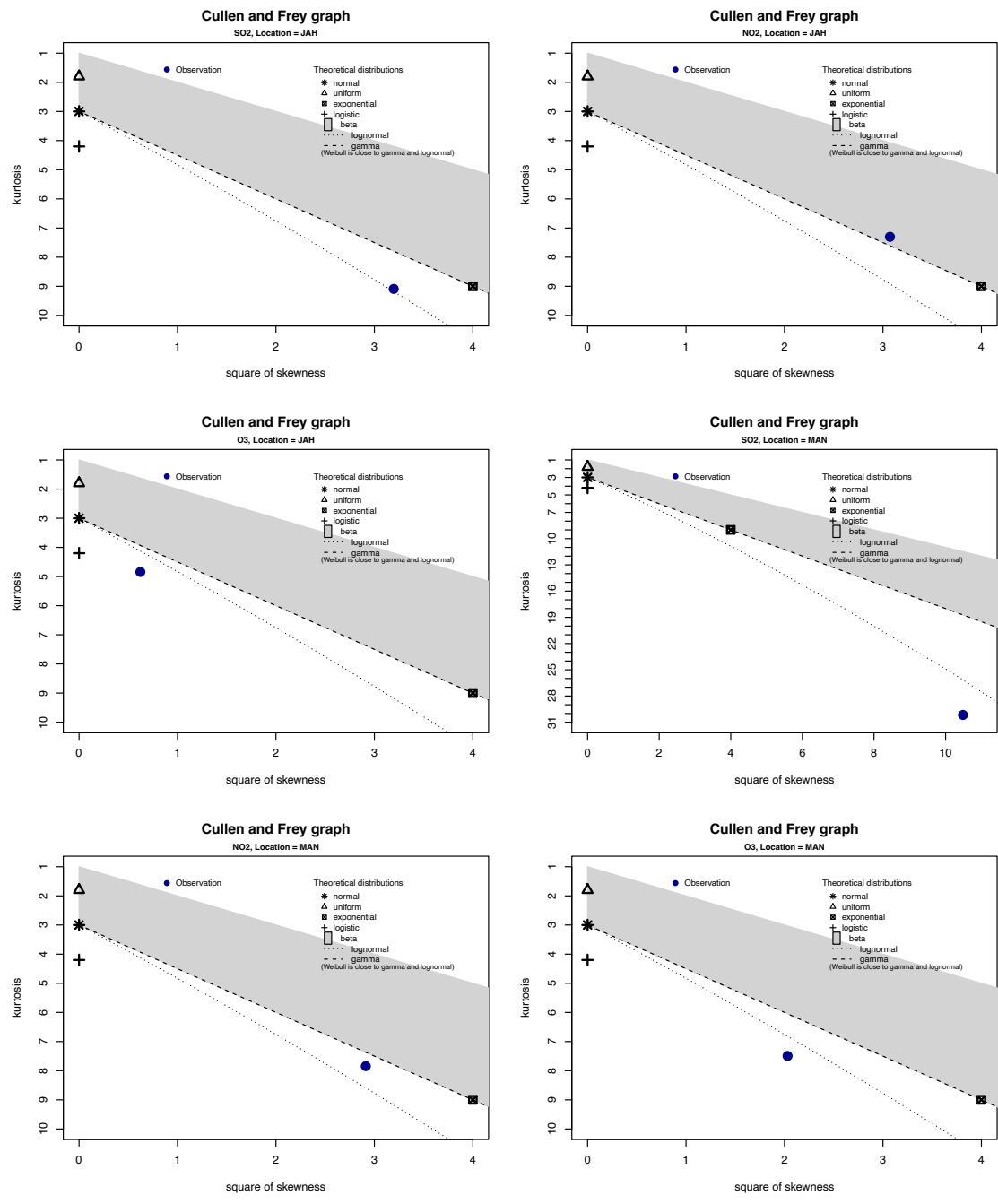


Figure 6.6: The distribution performance for Air quality index (AQI) using Cullen and Frey graphs for five governorate monitoring fixed stations in Kuwait during the period from 2013 to 2020. This figure captured the distribution performance for SO_2 , NO_2 and O_3 monitored by JAH and MAN fixed stations.

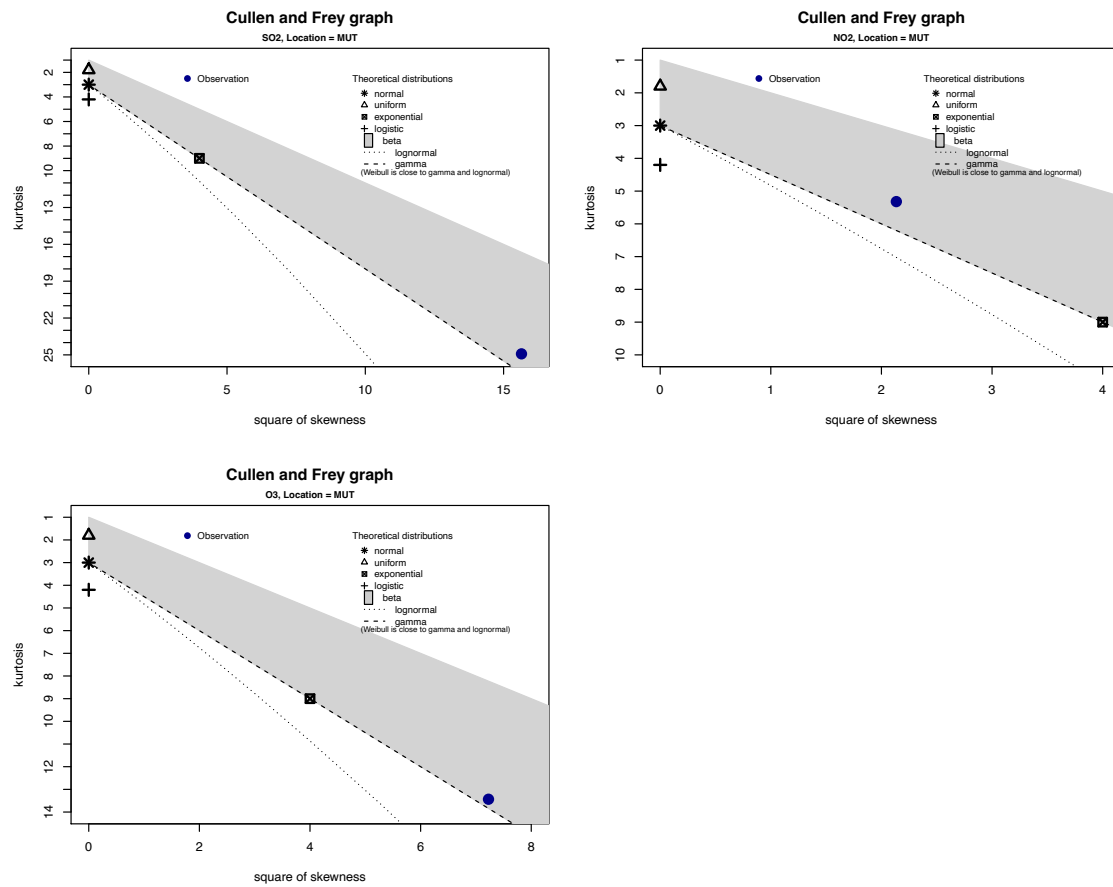


Figure 6.7: The distribution performance for Air quality index (AQI) using Cullen and Frey graphs for five governorate monitoring fixed stations in Kuwait during the period from 2013 to 2020. This figure captured the distribution performance for SO_2 , NO_2 and O_3 monitored by MUT fixed station.

Table 6.3: In-sample transformation efficacy measured by probability density function (Pdf) on the original samples ($n = 16,480$) after transformation using skewness results. Values close to one indicate normally transformed data.

Pollutants	Location/methods	ASA	FAH	JAH	MAN	MUT
SO_2	Arc Sin	8.190	5.985	3.929	40.996	17.436
	Box-Cox	6.511	3.331	2.257	44.015	12.472
	Lambert S	11.472	3.495	2.369	43.510	22.670
	Log Transform	7.863	6.552	4.181	42.530	13.743
	No Transform	79.329	19.566	19.142	55.899	83.378
	Square Root	24.815	4.362	4.828	46.269	32.833
	Yeo-Johnson	7.364	3.335	1.951	43.826	13.341
	Selected Transformation	Box-Cox	Box-Cox	Yeo-Johnson	Log Transform	Yeo-Johnson
NO_2	Arc Sin	1.433	2.483	9.362	2.365	1.959
	Box-Cox	1.351	1.876	5.550	2.330	2.539
	Lambert S	1.655	1.946	5.376	2.669	3.991
	Log Transform	1.422	2.524	9.441	2.440	2.124
	No Transform	17.057	9.349	26.214	17.736	26.204
	Square Root	5.496	2.851	8.629	5.927	8.223
	Yeo-Johnson	1.310	1.918	5.534	2.011	2.142
	Selected Transformation	Yeo-Johnson	Box-Cox	Lambert S	Yeo-Johnson	Log Transform
O_3	Arc Sin	3.489	4.869	9.021	4.980	8.381
	Box-Cox	0.750	1.639	2.578	3.372	7.815
	Lambert S	0.814	1.476	2.131	3.168	10.126
	Log Transform	3.594	5.091	9.300	5.021	8.474
	No Transform	5.178	6.678	4.288	8.829	35.745
	Square Root	0.949	1.739	2.399	3.401	16.286
	Yeo-Johnson	0.802	1.568	2.524	3.435	7.563
	Selected Transformation	Box-Cox	Lambert S	Lambert S	Lambert S	Yeo-Johnson

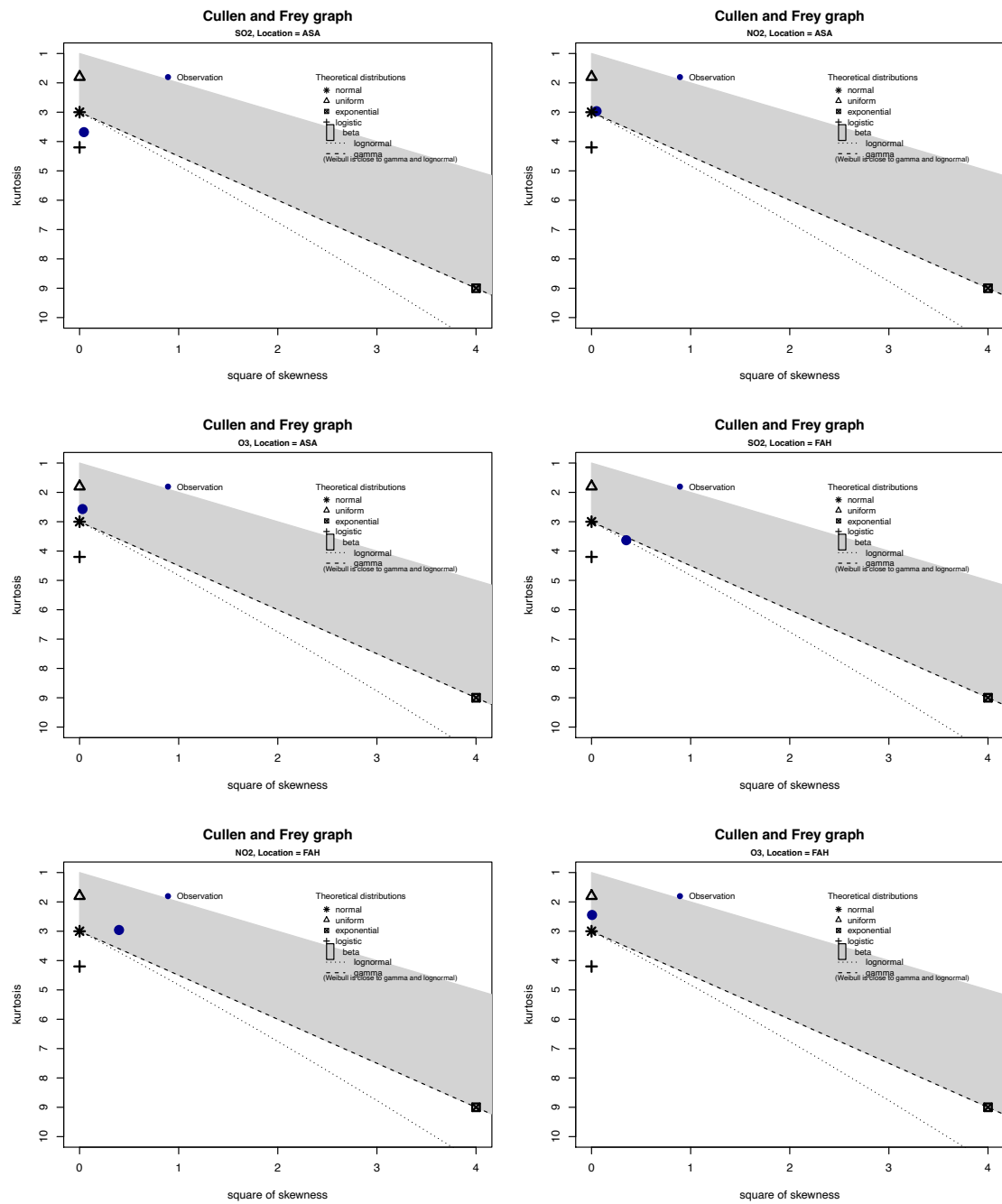


Figure 6.8: After implementing transformation, the distribution performance for Air quality index (AQI) using Cullen and Frey graphs for five governorate monitoring fixed stations in Kuwait during the period from 2013 to 2020. This figure captured the distribution performance for SO_2 , NO_2 and O_3 monitored by ASA and FAH fixed stations.

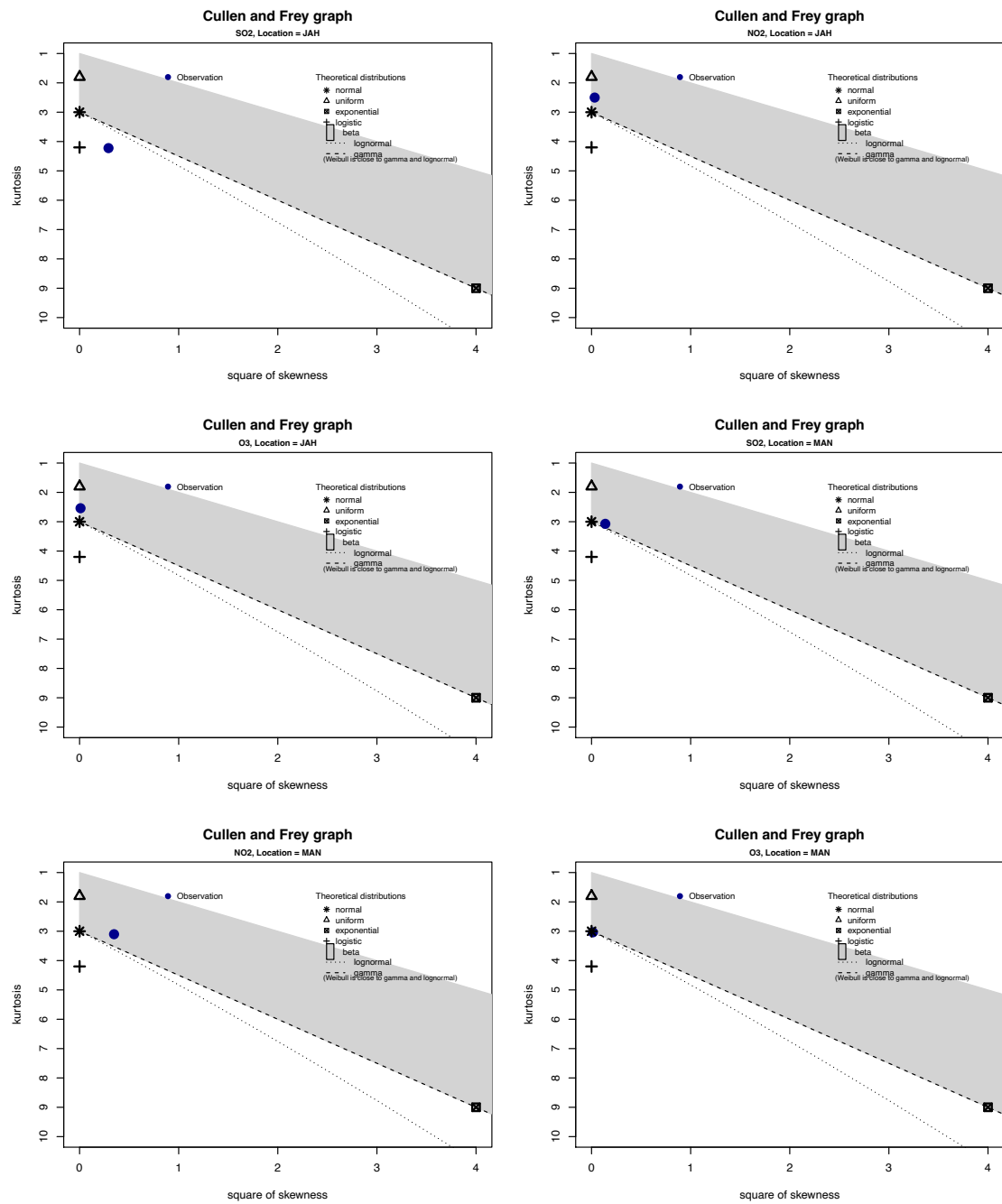


Figure 6.9: After implementing transformation, the distribution performance for Air quality index (AQI) using Cullen and Frey graphs for five governorate monitoring fixed stations in Kuwait during the period from 2013 to 2020. This figure captured the distribution performance for SO_2 , NO_2 and O_3 monitored by JAH and MAN fixed stations.

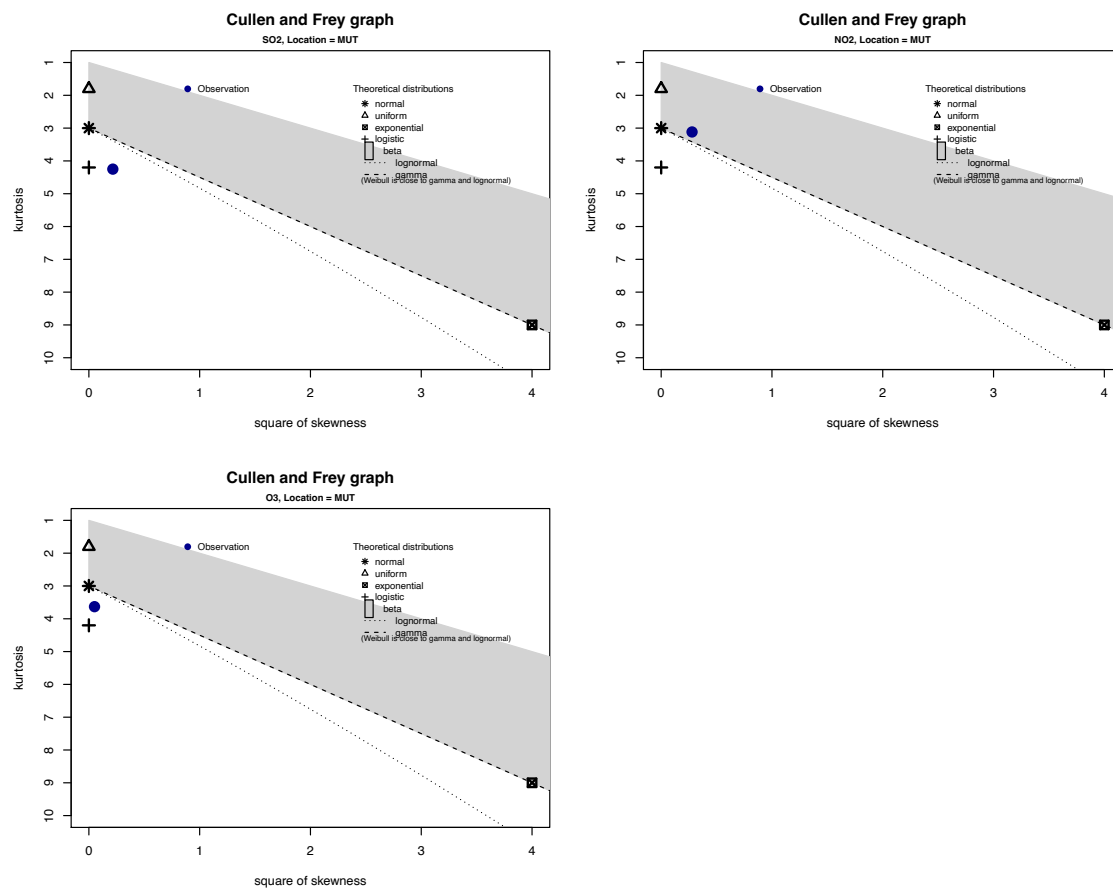
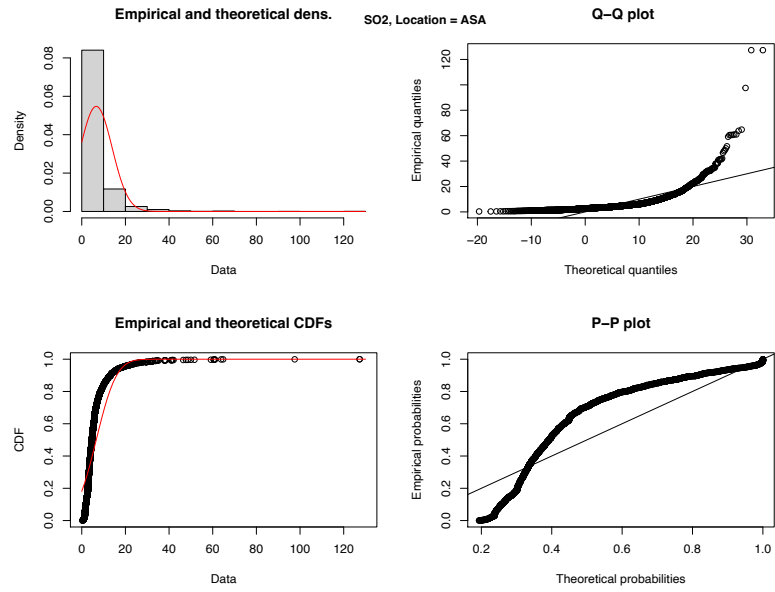
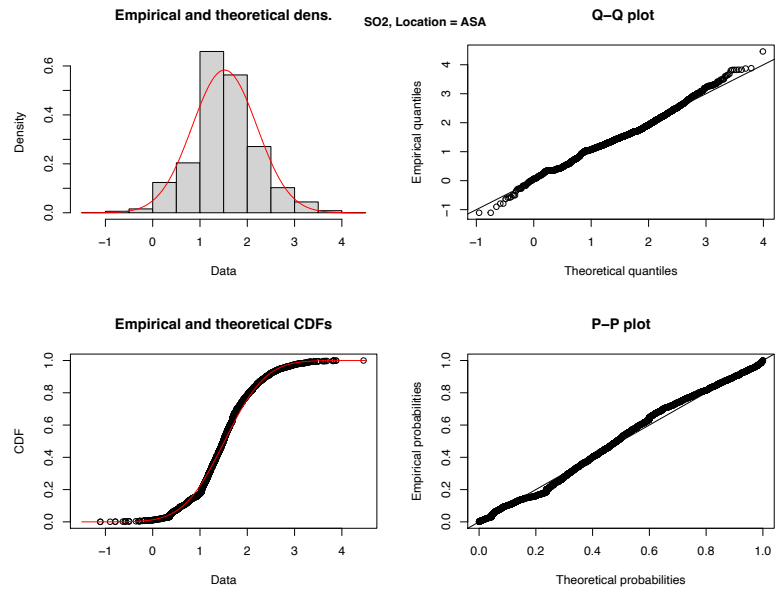


Figure 6.10: After implementing transformation, the distribution performance for Air quality index (AQI) using Cullen and Frey graphs for five governorate monitoring fixed stations in Kuwait during the period from 2013 to 2020. This figure captured the distribution performance for SO_2 , NO_2 and O_3 monitored by MUT fixed station.

The following figures (6.11, 6.12 and 6.13) show normality performance for SO_2 , NO_2 and O_3 from ASA monitoring fixed station before and after the Box-Cox and Yeo-Johnson transformation methods:

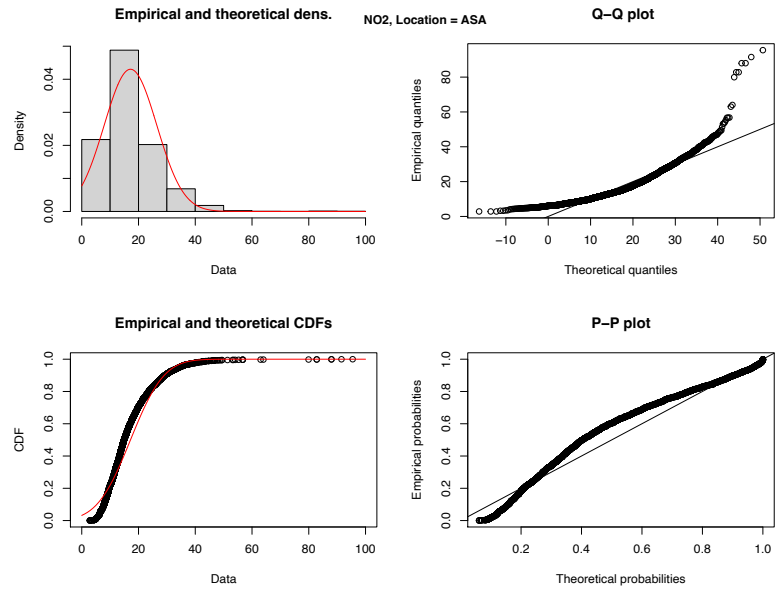


(a) SO_2 measured in ASA location before Box-Cox transformation

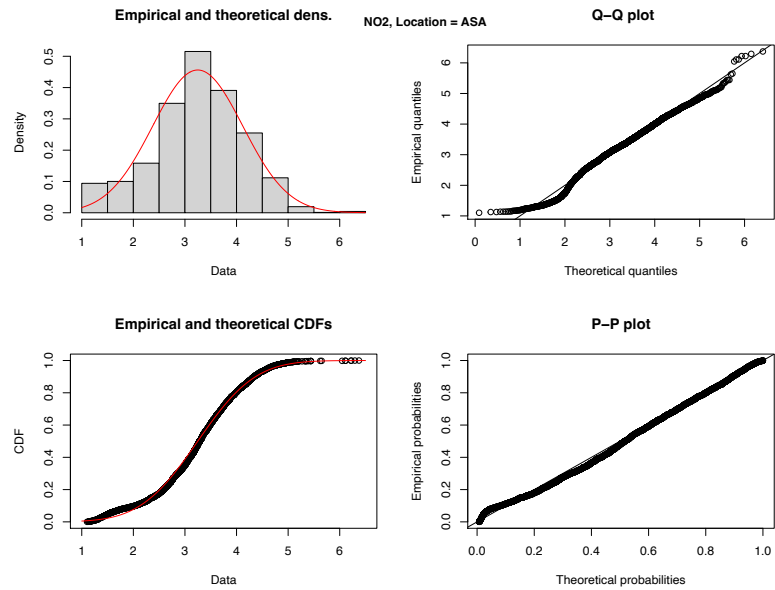


(b) SO_2 measured in ASA location after Box-Cox transformation

Figure 6.11: Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for SO_2 before and after the Box-Cox transformation; it is obvious that the Box-Cox transformation enhances the normality performance for SO_2 in ASA location.

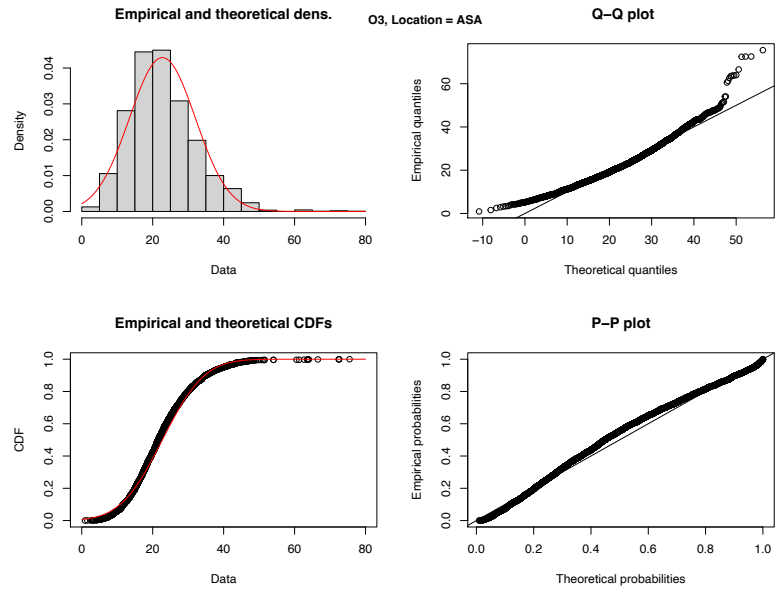


(a) NO_2 measured in ASA location before Yeo-Johnson transformation

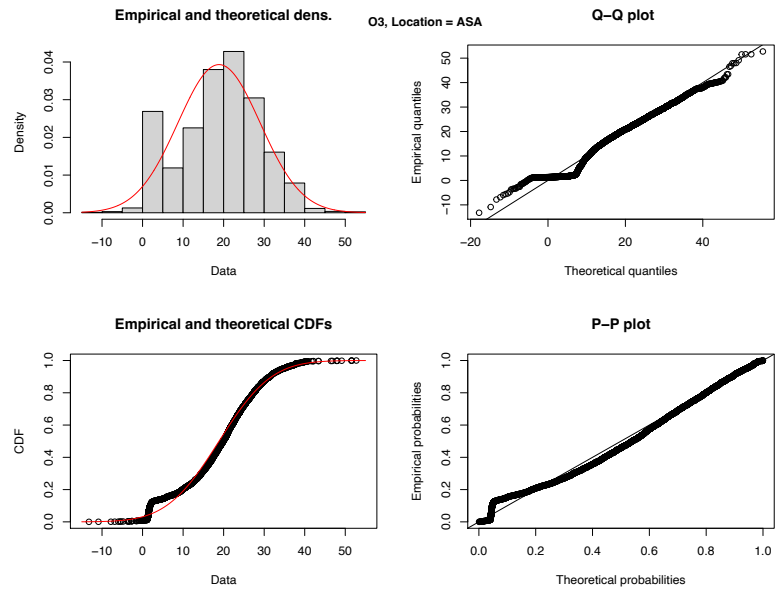


(b) NO_2 measured in ASA location after Yeo-Johnson transformation

Figure 6.12: Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for NO_2 before and after the Box-Cox transformation; it is obvious that the Yeo-Johnson transformation enhances the normality performance for NO_2 in ASA location.



(a) O_3 measured in ASA location before Box-Cox transformation



(b) O_3 measured in ASA location after Box-Cox transformation

Figure 6.13: Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for O_3 before and after the Box-Cox transformation; it is obvious that the Box-Cox transformation enhances the normality performance for O_3 in ASA location.

All other pollutants' distribution performance for the other monitoring fixed stations are attached in Appendix C [see figures from C.1 to C.12].

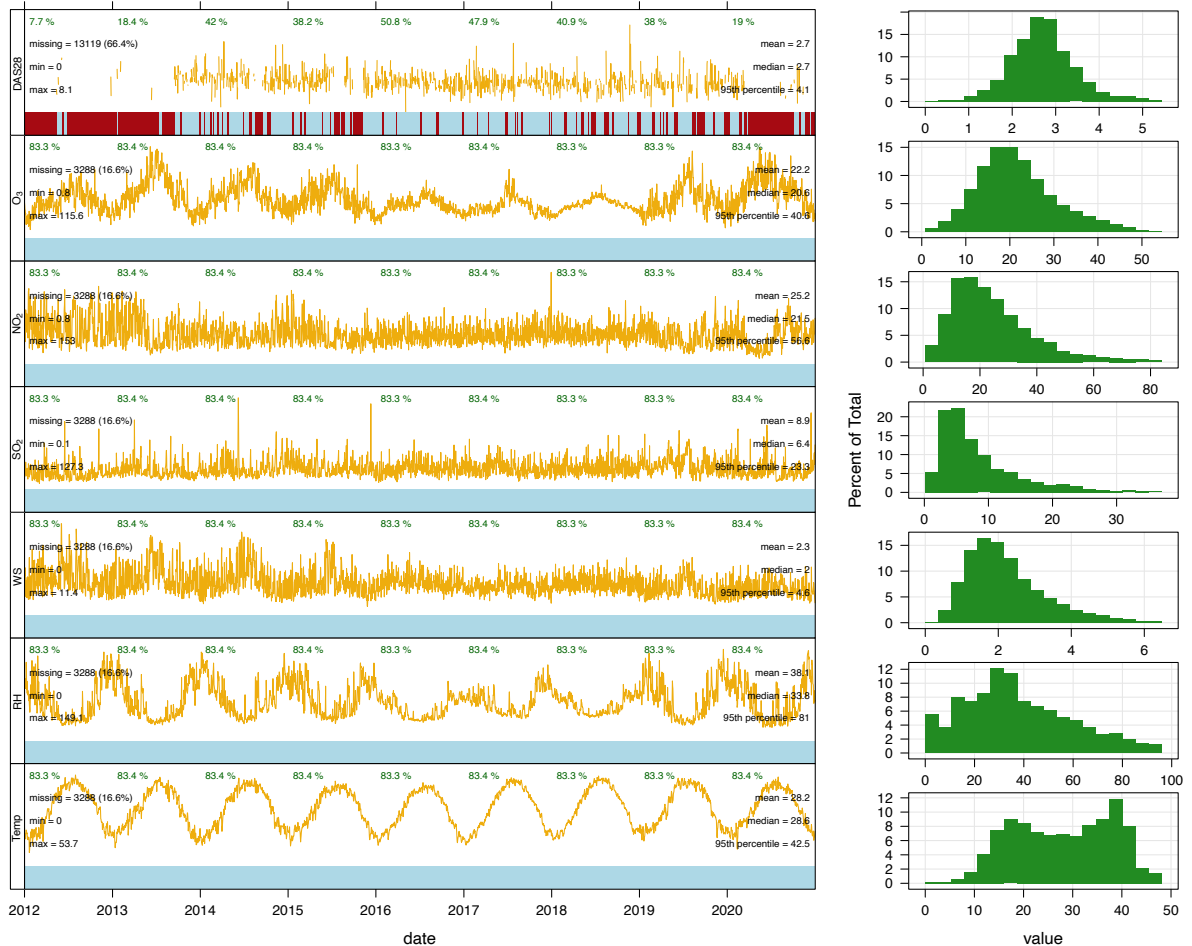


Figure 6.14: Descriptive statistics over time for each numerical variable in our data frame (DAS28, SO_2 , NO_2 , O_3 , Temp, RH and WS), a plot is made, shown in the left panel, showing where data exist (blue) and missing data (red). For clarity, only running sequences of ≥ 24 hours of missing.

In figure 6.14, the plots in the left panel demonstrate the time series data, where blue shows the presence of data and red missing data. The 24-hour mean values are also shown in pale yellow scaled to cover the range in the data from zero to the maximum daily

value. As such, the daily values do not necessarily convey quantitative information, but they indicate instead an overall trend. For each pollutant, the overall summary statistics are given. Moreover, the yearly captured percentage data is shown in green font. The distribution of each variable is presented using a histogram plot in the panels on the right. Therefore, it is obvious that the first and the last parts of the time series for DAS28 are missing. It is also evident that the time series stops at the end of 2020. Each panel shows statistical summaries, which include: number of missing points (with percentage shown in parentheses), minimum, maximum, mean, median and the 95th percentile. For each year, the data capture (%) is shown in green font. So, for example, the data capture for O_3 in 2012 was 83.4%, and in another example, the data capture for DAS28 in 2014 was 42%. The pale yellow line gives an indication of the variation in values over time expressed as a daily mean.

6.9 DAS28 OLS Models

A linear regression analysis was conducted to assess whether SO_2 , NO_2 , O_3 , Temperature, relative humidity (RH), and wind speed (WS) significantly predicted DAS28.

Table 6.4 presents the results of the linear regression model for each location. The results of the linear regression model were significant, $F(6, 1286) = 4.98, p < .001, R^2 = 0.066$, indicating that approximately 6.6% of the variance in DAS28 is explainable by SO_2 , NO_2 , O_3 , Temp, RH, and WS. In ASA and JAH stations, SO_2 did not significantly predict DAS28, p-value > 0.05 . Based on this sample, a one-unit increase in SO_2 does not have a significant effect on DAS28. However, our results showed that in MAN station, SO_2 significantly predict DAS28 as $B = 0.102, p = .018$. This indicates that on average, a one-unit increase of SO_2 will increase the value of DAS28 by 0.102 units. In ASA station, NO_2 significantly predicted DAS28, $B = 0.178, t(1286) = 2.51, p = .012$. This indicates that on average, a one-unit increase of NO_2 will increase the value of DAS28 by 0.178 units. In all stations except FAH, O_3 did not significantly predict DAS28, $B = 0.01, t(1286) = 0.82, p = .410$.

In ASA and FAH stations, temperature significantly predicted DAS28, $B = -0.007$, $t(1286) = -2.42$, $p = .015$ and $B = -0.012$, $t(1286) = -2.42$, $p = .015$ respectively. This indicates that on average, a one-unit increase of Temp will decrease the value of DAS28 by 0.012 or 0.007 units depending on the living address. Also, in MAN station, RH significantly predicted DAS28, $B = -0.006$, $t(1286) = -3.31$, $p < .001$. This indicates that on average, a one-unit increase of RH will decrease the value of DAS28 by 0.006 units. For ASA and FAH stations, RH will significantly decrease the value of DAS28 by 0.005 units for those patients living close to FAH and ASA stations. In MAN station, WS significantly predicted DAS28, as $B = 0.048$, $t(1286) = 2.13$, $p = .018$. This indicates that on average, a one-unit increase of WS will increase the value of DAS28 by 0.048 units.

As a result, the regression analysis models developed in this work can be used in an early warning system to aid in the mitigation and prevention of illness cases using predictor information. However, there are also some weaknesses in the regression performance as it shown in the R-square for the models. One of the study's flaws is that it only looks at the effects of air pollution factors on disease cases based on data collected during a specific time period. Conclusively, regression models will not be sufficient to explain RA disease activity score using the information of the air pollutants. We will present the multivariate time series approach to better explain and predict DAS28 using information of AQI for SO_2 , NO_2 and O_3 .

Table 6.4: OLS regression models of air pollution's impact on DAS28 in Kuwait during the period from 2012 to 2020.

	<i>Dependent variable:</i>			
	DAS28			
	(ASA)	(FAH)	(JAH)	(MAN)
SO_2	-0.042 (0.041)	-0.052* (0.020)	-0.045 (0.028)	0.102* (0.018)
NO_2	0.178* (0.071)	0.010 (0.028)	0.004* (0.002)	0.002 (0.040)
O_3	0.018 (0.017)	0.005 (0.003)	0.009** (0.003)	-0.002 (0.003)
Temp	-0.007* (0.003)	-0.012** (0.003)	0.002 (0.003)	-0.003 (0.003)
RH	-0.005** (0.001)	-0.005** (0.002)	0.002** (0.001)	-0.006* (0.002)
WS	0.041* (0.018)	-0.032 (0.031)	-0.004 (0.031)	0.048** (0.018)
Constant	2.530** (0.270)	3.387** (0.228)	2.480** (0.142)	2.682** (0.249)
Observations	1,328	1,330	1,331	1,330
R^2	0.066	0.048	0.064	0.052
Adjusted R^2	0.061	0.045	0.061	0.049

Note:

* $p < 0.05$; ** $p < 0.01$

6.10 Observations' Time Line

The monitored air pollution variables for NO_2 , SO_2 , and O_3 ; were saved based on hourly observations (24 records in a day for each pollutant) from the monitoring stations. Then, all pollutants' observations were aggregated according to a daily basis. The reason for this aggregation is because the RA disease activity scores (DAS28) values were collected based on daily observations, and that will make the databases more convenient to be matched. Also, there is another reason related to missing values' treatment, because daily aggregation will reduce the missing values' rate, and that will promote the quality level of the combined dataset.

Figure 6.15 presents the time series plots for the pollutant with the disease activity score (DAS28) for RA patients during the period from 2012 to 2020. The grey line is the time series for DAS28 and the red line is the time series for the pollutant. It is obvious in some plots in figure 6.15, that the time series for both pollutant and DAS28 are moving slightly around each other. For example, if we look at the figure 6.15 part (c), we can see that SO_2 and DAS28 are moving slightly around each other which gives an indication that there is a Granger causality between SO_2 and DAS28. With NO_2 in location ASA, we can see that there is a Granger causality between NO_2 and DAS28. So, figure 6.15 suggests that for some pollutants in some locations there is a causality association between the RA disease activity score with the exposure to the pollutant. To judge and confirm this long or short run relationship, we will use the cointegration test in section 6.13 and causality test which was explained in section 6.14.

This part of the analysis depicts each location's mean monthly pollutant readings along three percentiles: 50th, 5th, and 95th, to see if similar changes occur. Plotting these numbers over time reveals variations in the data's low, mid, and high ranges. To show monthly and long-term fluctuation, these lines have been smoothed. The plots also provide a 95% confidence interval around the long-term trend line between the pollutant and DAS28, which is indicated by shading around the line. Figures 6.16 to 6.18 explain the smooth time series model in order to show if there is any possible trend between the

pollutant and DAS28 by location.

As illustrated in figure 6.16, the monthly mean SO_2 concentrations as measured at the four monitoring sites (ASA, FAH, MAN and JAH), went through different relationships during the considered time interval. A downward trend in SO_2 concentrations was observed between the years 2012 and 2020 in JAH and MAN sites associated with a downward trend for DAS28 for example.

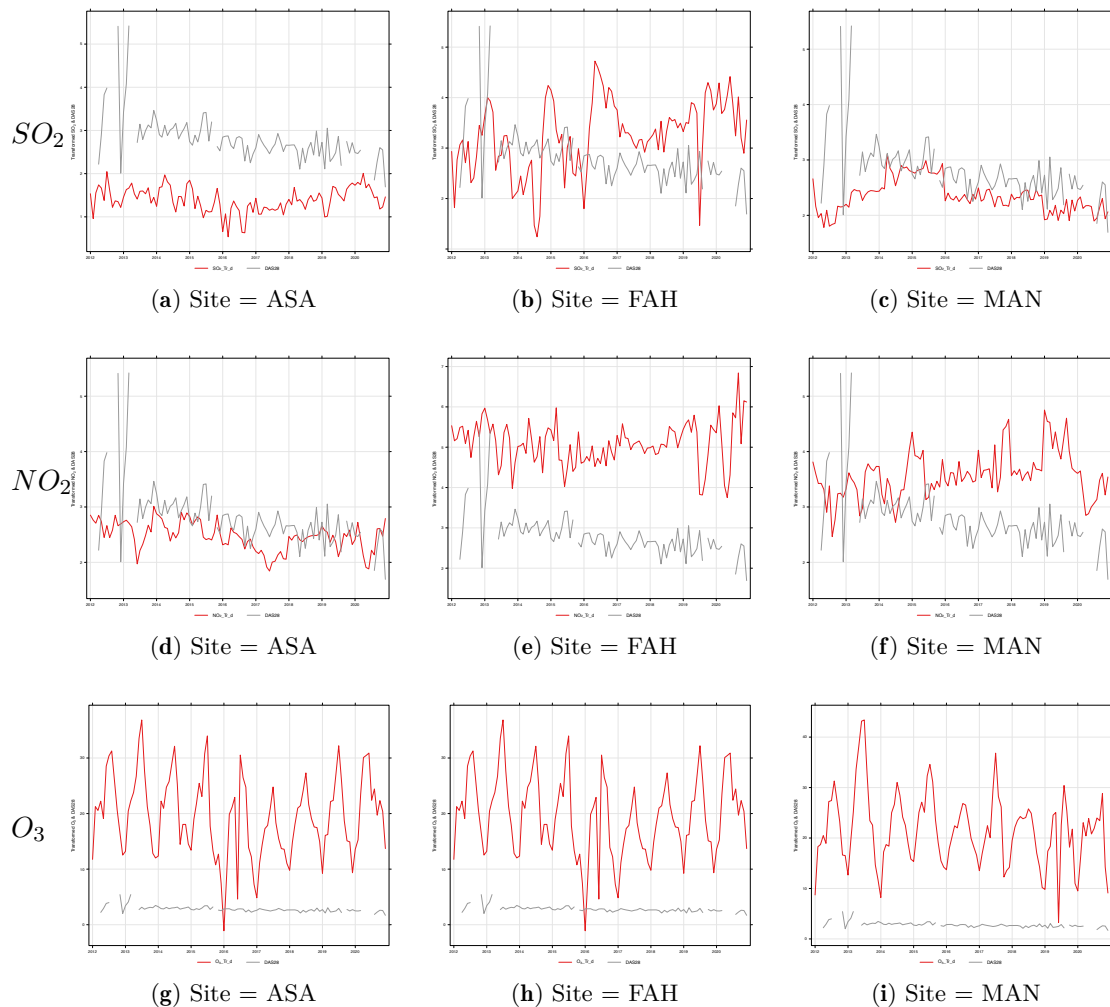


Figure 6.15: Multivariate time series graphs between DAS28 (grey line) with pollutant concentration line (red line) from the period from 2012 to 2020 based on monitoring station name.

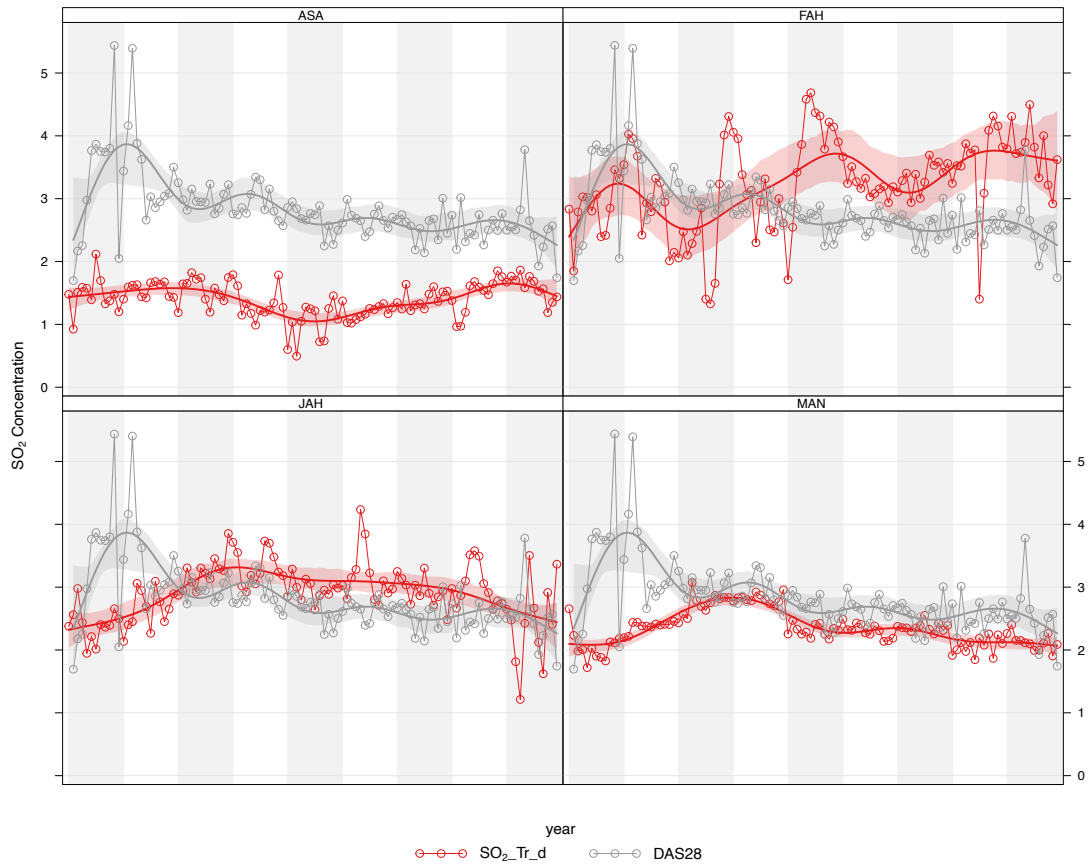


Figure 6.16: Long-term (2012-2020) trends of SO_2 concentrations and DAS28 calculated using the smooth trend method based on the mean measurements for four locations.

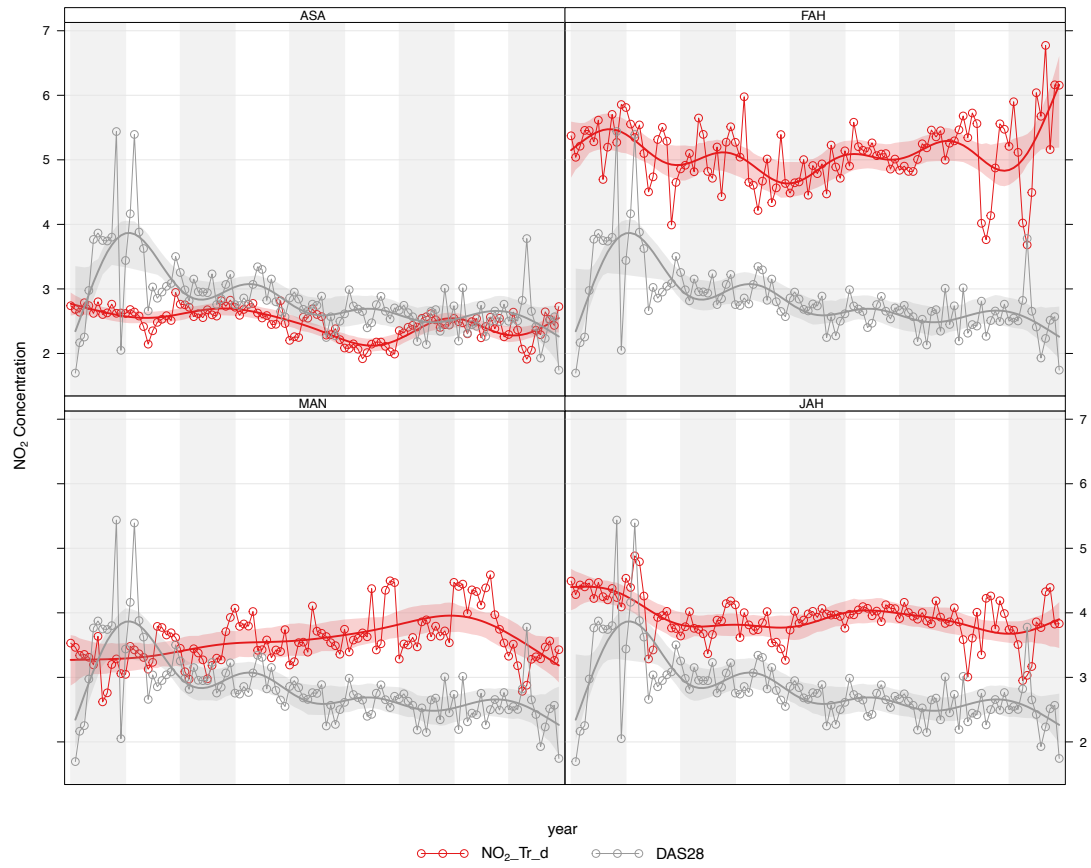


Figure 6.17: Long-term (2012-2020) trends of NO_2 concentrations and DAS28 calculated using the smooth trend method based on the mean measurements for four locations.

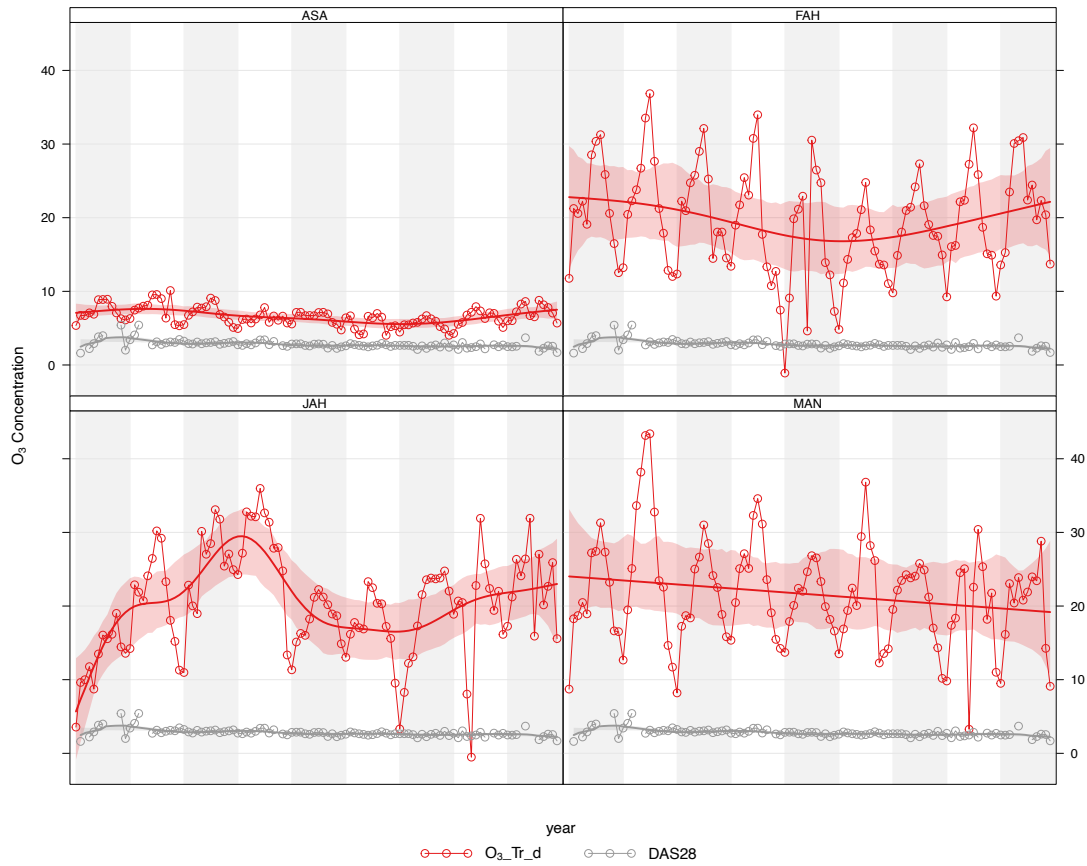


Figure 6.18: Long-term (2012-2020) trends of O_3 concentrations and DAS28 calculated using the smooth trend method based on the mean measurements for four locations.

6.11 Stationarity Test

6.11.1 Unit Root Test

Because the data in this study contains all time series, a unit root test for each variable in all study locations is required before the main analysis in order to avoid misleading or spurious regression.

The KPSS test (Sephton, 1995) and Augmented Dicky-Fuller (1981) (ADF) test are used to examine unit root properties of the series. The ADF test is founded on the null

hypothesis (H_0) that the levels of the series have a unit root. The alternative hypothesis (H_1) states that the levels of the series are stationary (Dickey and Fuller, 1981).

6.11.2 Augmented Dicky-Fuller (ADF) Test

Table 6.5 shows the results of the Augment Dickey-Fuller Unit Root Test (ADF), and it shows that O_3 , SO_2 , NO_2 and $\log(DAS28)$ are stationary at first difference, which is known as $I(1)$. This result fulfilled the prerequisite of the VECM model, that none of the data or variables is 2nd difference stationary or $I(2)$. The ADF estimations show that at the first difference, all variables are stationary. This fulfills the basic requirement of cointegration.

Table 6.5: ADF root test with constant and with constant and trend for EPA datasets (ASA, FAH, MAN and JAH monitoring fixed stations).

Station	Type Variables	with constant		with constant and trend	
		level t-statistic	First difference t-statistic	level t-statistic	First difference t-statistic
ASA	SO_2	-10.413***	-22.604***	-10.414***	-22.600***
	NO_2	-6.242***	-22.453***	-6.573***	-22.451***
	O_3	-7.771***	-28.983***	-7.939***	-28.978***
	$\log(DAS28)$	-5.505***	-22.342***	-5.525***	-22.341***
FAH	SO_2	-9.973***	-22.067***	-10.596***	-22.064***
	NO_2	-12.764***	-22.488***	-12.873***	-22.487***
	O_3	-6.320***	-25.976***	-6.346***	-25.974***
	$\log(DAS28)$	-5.505***	-22.342***	-5.525***	-22.341***
MAN	SO_2	-6.636***	-23.343***	-8.293***	-23.339***
	NO_2	-8.045***	-21.274***	-8.172***	-21.272***
	O_3	-7.263***	-31.094***	-7.585***	-31.093***
	$\log(DAS28)$	-5.505***	-22.342***	-5.525***	-22.341***
JAH	SO_2	-11.150***	-24.928***	-11.306***	-24.923***
	NO_2	-8.460***	-22.628***	-9.583***	-22.624***
	O_3	-7.677***	-22.764***	-7.852***	-22.764***
	$\log(DAS28)$	-5.505***	-22.342***	-5.525***	-22.341***

* $p < .05$, ** $p < .01$, *** $p < .001$

6.11.3 Augmented Dickey-Fuller Generalised Least Squares (ADF-GLS) Unit Root Test Results

The results of the ADF-GLS unit roots test for the level and first difference data are shown in Table 6.6. The ADF-GLS test confirms that all variables used in the research [SO_2 , NO_2 , O_3 and $\log(DAS28)$] in all four regions are integrated of first order $I(1)$, as proposed by Elliott et al. (1992) and Dickey and Fuller (1979b). As a result, the null hypothesis of the ADF-GLS test is non-stationarity, implying that the null hypothesis must be rejected before proceeding. The ADF-GLS test is used to establish the order of integration for each variable, which is a key step in our cointegration investigation. The cointegration test and the Error Correction Model (ECM) can be employed if all variables are determined as integrated of first order using the ADF-GLS test. We discovered that all series are integrated of first order in table 6.6. As a result, the analysis rejects the null hypothesis that the variables have a unit root, implying that they are non-stationary, and concludes that the undifferenced data is stationary.

6.11.4 KPSS Root Test Results

The ADF is criticised (Loganathan and Subramaniam, 2010) for having low power as compared to the KPSS test. The KPSS model is employed as a confirmatory test to the ADF test in this study. The KPSS test pits the null hypothesis that the series variables have no unit root against the alternative hypothesis that the series variables do have a unit root.

The KPSS test has the hypothesis $H_0 : \delta = 0$, there is no root unit (stationary data), and the opponent is $H_1 : \delta < 0$, there is unit root (data not stationary). Test results of SO_2 , NO_2 , O_3 and $\log(DAS28)$ using the KPSS unit root test can be found in table 6.7.

Using the KPSS test, each series is first difference stationary at the 1%, 5% and 10% levels, according to the numbers in table 6.7. As a result, we use the KPSS test result to conduct a cointegration test among all stationary series of the same order, implying that the SO_2 , NO_2 , O_3 and $\log(DAS28)$ series are stationary at their first differences [they are $I(1)$ integrated]. Because the p-value was more than 0.05 (the null and alternate

Table 6.6: AD-GLS root test with constant and with constant and trend for EPA datasets (ASA, FAH, MAN and JAH monitoring fixed stations).

Station	Type Variables	constant		constant and trend	
		level t-statistic	First difference t-statistic	level t-statistic	First difference t-statistic
ASA	SO_2	-6.47075***	-1.96593	-6.59976***	-4.02148**
	NO_2	-3.42669**	-1.11922	-4.23287***	-2.44038
	O_3	-3.65971***	-1.29717	-5.0391***	-2.79732
	$\log(DAS28)$	-2.42743*	-75.2977***	-3.61741**	-2.7367
FAH	SO_2	-3.74548***	-1.63613	-4.08276**	-2.17749
	NO_2	-3.60929***	-74.3357***	-4.57909***	-6.34743***
	O_3	-4.30975***	-0.85265	-4.33135***	-2.07934
	$\log(DAS28)$	-2.42743*	-75.2977***	-3.61741**	-2.7367
MAN	SO_2	-3.64533***	-87.089***	-3.77802**	-3.74617**
	NO_2	-3.1167**	-2.20553*	-4.77832***	-4.49956***
	O_3	-2.83155**	-2.52982*	-3.42937**	-5.15543***
	$\log(DAS28)$	-2.42743*	-75.2977***	-3.61741**	-2.7367
JAH	SO_2	-2.51654*	-84.2394***	-3.97833**	-84.3189***
	NO_2	-2.6942**	-0.751061	-5.1681***	-1.97689
	O_3	-3.53047***	-5.97657***	-4.1241**	-5.68603***
	$\log(DAS28)$	-2.42743*	-75.2977***	-3.61741**	-2.7367

* $p < .05$, ** $p < .01$, *** $p < .001$, in ADF-GLS test indicate the rejection of the null hypothesis that the series has a unit root at 1%, 5% and 10% levels of significance. The optimum lag of 7 was determined using SBC.

hypotheses for the KPSS test are the opposite of those of the ADF test), the KPSS results for the SO_2 , NO_2 , O_3 and $\log(DAS28)$ suggest the presence of a unit root while the data is stationary, $p > 0.05$ except the KPSS results for the SO_2 related to MAN station.

Table 6.7: KPSS test with constant and with constant and trend for EPA datasets (ASA, FAH, MAN and JAH monitoring fixed stations).

Station	Type Variables	constant		constant and trend	
		level t-statistic	First difference t-statistic	level t-statistic	First difference t-statistic
ASA	SO_2	1.25242**	0.005	1.23984**	0.004
	NO_2	2.54604**	0.002	0.588188**	0.007
	O_3	1.57345**	0.007	0.880004**	0.880
	$\log(DAS28)$	1.19348***	0.008	1.16369***	0.008
FAH	SO_2	2.08946**	0.017	0.194559*	0.007
	NO_2	0.497695*	0.015	0.228419**	0.006
	O_3	0.493232*	0.011	0.44757**	0.010
	$\log(DAS28)$	1.19348**	0.008	1.16369**	0.008
MAN	SO_2	6.4743**	0.007	0.758877**	0.006
	NO_2	1.21025**	0.006	0.447266**	0.005
	O_3	0.938624**	0.024	0.073	0.011
	$\log(DAS28)$	1.19348**	0.008	1.16369**	0.008
JAH	SO_2	1.21666**	0.008	0.684289**	0.008
	NO_2	4.43434**	0.011	0.338869**	0.006
	O_3	1.12034**	0.015	0.482687**	0.009
	$\log(DAS28)$	1.19348**	0.008	1.16369**	0.008

* $p < .05$, ** $p < .01$, *** $p < .001$, *, **, *** denotes rejection of the null hypothesis of trend stationarity at the 5%, 1%, and 0.1% significance levels, respectively.

6.11.5 Phillips-Perron (PP) Test Results

The Philips-Perron Unit Root test was used to determine the series' stationary level. According to the results in table 6.8, for the Phillips-Perron test (PP) for SO_2 , NO_2 , O_3 and $\log(DAS28)$, the p-values for all of the sample variables were close to zero (i.e. $p < 0.05$). The results showed that all of the sample variables' data had reached stationarity in both the level base and the first difference base according to the Phillips-Perron test (Phillips and Perron, 1988).

Table 6.8: Phillips-Perron test with constant and with constant and trend for EPA datasets (ASA, FAH, MAN and JAH monitoring fixed stations).

Station	Type Variables	constant		constant and trend	
		level	First difference	level	First difference
ASA	SO_2	-43.21843***	-172.6618***	-43.21523***	-172.6233***
	NO_2	-34.75011***	-153.8485***	-35.86924***	-153.8465***
	O_3	-22.4985***	-114.77***	-22.98516***	-114.7429***
	$\log(DAS28)$	-45.00959***	-205.7314***	-44.99826***	-205.6868***
FAH	SO_2	-42.56249***	-168.2883***	-43.00987***	-168.2655***
	NO_2	-39.09392***	-151.3224***	-39.09981***	-151.3189***
	O_3	-21.58105***	-118.4749***	-21.61616***	-118.4669***
	$\log(DAS28)$	-45.00959***	-205.7314***	-44.99826***	-205.6868***
MAN	SO_2	-43.062***	-193.837***	-47.34359***	-193.7997***
	NO_2	-40.13827***	-166.533***	-40.40814***	-166.4973***
	O_3	-18.7203***	-107.402***	-19.3696***	-107.4398***
	$\log(DAS28)$	-45.00959***	-205.7314***	-44.99826***	-205.6868***
JAH	SO_2	-45.04141***	-182.8147***	-45.08749***	-182.7723***
	NO_2	-41.72624***	-172.1351***	-43.69968***	-172.1097***
	O_3	-21.64522***	-107.6285***	-21.94752***	-107.6275***
	$\log(DAS28)$	-45.00959***	-205.7314***	-44.99826***	-205.6868***

* $p < .05$, ** $p < .01$, *** $p < .001$

6.12 Lag Selection Criteria

Prior to performing the Johansen cointegration test, variables were entered as levels into a VAR to determine the optimal number of lags needed in the cointegration analysis. In addition to the likelihood ratio (LR) test, three criteria were used to identify the ideal lag length: Bayesian Schwartz Information Criteria (BIC), Akaike Information Criteria (AIC), and Hannan-Quinn Criteria (HQC). Based on HQC and BIC criteria, we chose a lag of the 7th order based on assessment of different models. Table 6.9 present the lag selection for ASA location. We can see from table 6.9, according to BIC, lag 7 has the lowest BIC among the other 20 lags. These results agreed with those of the other locations (FAH, MAN and JAH).

Table 6.9: VAR Lag order selection - ASA Station.

lags	loglik	p(LR)	AIC	BIC	HQC
1	-13999.904		9.835909	10.395764	10.037628
2	-13820.482	0.000	9.723282	10.316069	9.936866
3	-13715.614	0.000	9.662036	10.287756	9.887486
4	-13639.448	0.000	9.620571	10.279223	9.857887
5	-13608.654	0.000	9.610375	10.30196	9.859557
6	-13531.047	0.000	9.567917	10.292435	9.828965
7	-13360.841	0.000	9.461641	10.219091*	9.734555*
8	-13345.477	0.015	9.462079	10.252462	9.746858
9	-13320.167	0.000	9.455663	10.278978	9.752308
10	-13289.215	0.000	9.445358	10.301607	9.753869
11	-13281.546	0.500	9.4511	10.340281	9.771477
12	-13262.490	0.001	9.448994	10.371108	9.781237
13	-13225.729	0.000	9.434686	10.389732	9.778794
14	-13161.564	0.000	9.401491*	10.38947	9.757465
15	-13156.184	0.824	9.408811	10.429722	9.776651
16	-13139.939	0.009	9.408642	10.462486	9.788348
17	-13121.667	0.002	9.407076	10.493853	9.798648
18	-13117.692	0.950	9.415363	10.535073	9.818801
19	-13106.261	0.117	9.418512	10.571154	9.833815
20	-13087.083	0.001	9.416322	10.601896	9.843491

* indicates lag order selected by the criterion, and LR test statistic (each test at 5% level)

6.13 Johansen Cointegration Test

The Engle-Granger (Engle et al., 1987) methods can be used to assess the existence of cointegration between $\log(DAS28)$ and SO_2 , NO_2 and O_3 , after calculating the integration level for all the variables involved. The cointegration vector must be at level 1 to support the premise that air pollutants (SO_2 , NO_2 and O_3) and $\log(DAS28)$ are cointegrated.

The Johansen cointegration test was employed to see if there was a cointegration relationship between our variables. Most researchers prefer the Johansen cointegration test because it has the advantage of evaluating and estimating several long-run equilibrium relationships, which overcomes the limitations of single-equation approaches based

on restricted assumptions.

The outcomes of the Johansen cointegration test for the distinct stations are shown in table 6.10. The presence of three cointegration vectors is indicated by both the L-max and the Trace tests, however the L-max test marginally rejects the existence of a fourth cointegrating vector. This is confirmed also for all stations individually.

Table 6.10: Johansen Cointegration Test for the different monitoring stations.

Station	Rank	Eigenvalue	Trace test	p-value	Lmax test	p-value
ASA	0	0.077933	568.3	[0.0000]	236.51	[0.0000]
	1	0.059775	331.79	[0.0000]	179.67	[0.0000]
	2	0.028315	152.12	[0.0000]	83.73	[0.0000]
	3	0.023188	68.388	[0.0000]	68.388	[0.0000]
FAH	0	0.087122	576.84	[0.0000]	265.71	[0.0000]
	1	0.04579	311.13	[0.0000]	136.63	[0.0000]
	2	0.041648	174.5	[0.0000]	124	[0.0000]
	3	0.017175	50.5	[0.0000]	50.5	[0.0000]
MAN	0	0.10863	651.27	[0.0000]	335.21	[0.0000]
	1	0.052898	316.05	[0.0000]	158.43	[0.0000]
	2	0.029995	157.63	[0.0000]	88.774	[0.0000]
	3	0.023344	68.855	[0.0000]	68.855	[0.0000]
JAH	0	0.056878	523.43	[0.0000]	170.7	[0.0000]
	1	0.052237	352.73	[0.0000]	156.39	[0.0000]
	2	0.035548	196.34	[0.0000]	105.51	[0.0000]
	3	0.030679	90.831	[0.0000]	90.831	[0.0000]

The Johansen cointegration's trace statistics (Trace test) and maximum statistics (Lmax) were both used. The following equation mathematically represents the maximum statistics:

$$\text{Max statistic} = -S \ln(1 - L_{r+1}). \quad (6.4)$$

where S is the sample size and L is the i th largest canonical correlation. The null hypothesis of r cointegration is evaluated against the alternative hypothesis of $r+1$ using maximal statistics. The Johansen cointegration's maximum statistics illustrate that we reject the null hypothesis when the maximum statistic is greater than the critical value

at the 5% significance level under none, at most $r=1$, at most $r=2$, and at most $r=3$ cointegration.

In addition, the trace statistic tests the null hypothesis that the cointegration rank is equal or less than k versus the alternative that it is greater than k . The log-likelihood ratio, as estimated, is used to compute this trace test in the following equation:

$$\text{Ln} [L_{\max}(r)/L_{\max}(r + 1)]. \quad (6.5)$$

The null hypothesis that the cointegration rank is equal to r is rejected when the trace is bigger than the critical value for a certain rank. So, for our case, when r is greater or equal to 4, then the trace is lower than the critical value, and then we reject all cointegration rank greater or equal 4.

It can be deduced from the Johansen cointegration test that the null hypothesis of no cointegration is rejected for rank of zero at the 5% significance level for d (maximum eigenvalue tests and trace). We established that Lmax and trace tests both show 3 cointegrating equations based on the 5% level of significance (see table 6.10).

6.14 Causality Test

Even though the cointegration test can determine whether X and Y have a long-term equilibrium association, more research is needed to determine whether this link is causative. Regression of Y with both X and Y past values is better and more convincing than regression of Y with only Y past values. X is either the Granger or non-Granger cause of Y in the first scenario. In Chapter 5, under equation (5.48), the Granger test's form was described.

When computing multivariate time series models, one of the advantages is the ability to apply causality tests, which we could not do when examining univariate time series models. The result variable is determined by the cause variable, and changes in the cause variable lead to variations in the outcome variable. According to the formula (5.48), the outcomes are illustrated in Table 6.11.

Table 6.11 depicts that, in cases where the significance level is p-value ≤ 0.05 , the hypothesis, X (i.e. SO_2 , NO_2 or O_3) is not the Granger-cause of Y (i.e. $\log(DAS28)$), is rejected during lag phases (lag = 7), which means that air pollution is the Granger cause of disease activity score for RA patients.

Table 6.11 shows the results of the Granger causality test. At the 5% significance level, SO_2 , NO_2 or O_3 Granger-cause $\log(DAS28)$ for VAR(1), and, SO_2 and NO_2 did not Granger-cause $\log(DAS28)$ only in the FAH location. However, SO_2 , NO_2 and O_3 Granger-cause $\log(DAS28)$ for both the ASA and MAN locations. In the JAH location, SO_2 and O_3 Granger-cause $\log(DAS28)$.

According to the cointegration test, there is a long-term and stable cointegration relationship between the environmental pollution variables (SO_2 , NO_2 or O_3) and $\log(DAS28)$. Based on the reported F- statistics and P-values, we reject our null hypothesis for the regressions shown in Table 6.11. In the long term, we find a unidirectional causality from pollution variables (SO_2 , NO_2 or O_3) on $\log(DAS28)$.

Table 6.11: Granger causality test - Long-run Estimation Results.

Location	Hypothesis	Constant	Constant + Trend	Granger Cause
		F-Statistic	F-Statistic	
ASA	Constant	-3.826***	-2.540**	
	Does SO_2 Granger cause $\log(DAS28)$	-2.558**	-2.337**	Yes
	Does NO_2 Granger cause $\log(DAS28)$	3.577***	2.912***	Yes
	Does O_3 Granger cause $\log(DAS28)$	-8.800***	-8.926***	Yes
FAH	Constant	-3.725***	-3.383***	
	Does SO_2 Granger cause $\log(DAS28)$	-0.9534	-0.7065	No
	Does NO_2 Granger cause $\log(DAS28)$	-0.4586	-0.4818	No
	Does O_3 Granger cause $\log(DAS28)$	-8.944***	-8.975***	Yes
MAN	Constant	-10.30***	-8.416***	
	Does SO_2 Granger cause $\log(DAS28)$	3.619***	2.649***	Yes
	Does NO_2 Granger cause $\log(DAS28)$	5.754***	5.881***	Yes
	Does O_3 Granger cause $\log(DAS28)$	-5.247***	-5.406***	Yes
JAH	Constant	-12.12***	-9.562***	
	Does SO_2 does Granger cause $\log(DAS28)$	4.596***	4.499***	Yes
	Does NO_2 Granger cause $\log(DAS28)$	1.284	0.9439	No
	Does O_3 Granger cause $\log(DAS28)$	-4.676***	-4.731***	Yes

* $p < .05$, ** $p < .01$, *** $p < .001$

6.15 VAR Modelling Results

As previously said, the main goal of this chapter is to look into how air pollution, and disease activity score (DAS28) among RA patients are all interrelated in the State of

Kuwait and how they affect each other. A vector autoregression (VAR) model is used to capture these dynamics. The primary benefit of using a VAR model to study dynamics is that it regards all variables as endogenous, allowing for temporal analysis in an a theoretical framework. Every variable needs an equation clarifying its development based on its own lags and the lags of the other variables in the model, so all variables are treated symmetrically in a structural sense. VAR models require minimum understanding of the forces that influence each of the variables because all variables are viewed as endogenous. The following are the components of the VAR model:

$$Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \dots + \Phi_p Y_{t-p} + H + \varepsilon_t, \quad (6.6)$$

where p is the maximum lag with a nonzero coefficient matrix,

$$Y_t = (\text{DAS28}_t, \text{SO}_{2,t}, \text{NO}_{2,t}, \text{O}_{3,t})' \quad (6.7)$$

and $\Phi_i (i = 1, 2, \dots, p)$ is a 4×4 matrix of coefficients, H is a column vector of constants, ε_t is the error, which is considered as white noise. We allow for seven lags ($p = 7$) as recommended by the AIC and BIC information criteria.

The results in table 6.12 show the final VAR estimates to explain the relationship between $\log(\text{DAS28})$ with other air pollutants factors (SO_2 , NO_2 and O_3), in addition to weather factors which are temperature (Temp), relative humidity (RH) and wind speed (WS). There are three main findings deduced from table 6.12.

First, consider the direction of the relationship between $\log(\text{DAS28})$ and SO_2 . The results in table 6.12 explains the relationship between $\log(\text{DAS28})$ and SO_2 in the four locations, and we found a significant relationship between $\log(\text{DAS28})$ and SO_2 only in the MAN and JAH locations with positive impact equal to 1.905 and 2.118 respectively at lag 2. This means that after one day of the emission increasing from SO_2 , the disease activity score for RA patients living in MAN and JAH will increase by 1.905 and 2.118 respectively.

The second finding is about the direction of the relationship between $\log(\text{DAS28})$ and

NO_2 . The results in table 6.12 explain the relationship between $\log(DAS28)$ and NO_2 in the four locations as well, and we found a significant relationship between $\log(DAS28)$ and NO_2 only in the ASA and MAN locations with negative impact equal to -2.089 and -1.894 respectively at lag 2, however lag 5 for the JAH location has negative impact equal to -1.704. These results confirmed that the RA patients in ASA, MAN and JAH locations do not suffer by the increases of NO_2 in their locations and the reason is the major sources for NO_2 emissions are far away from their location. But, for the RA patients living in MAN or close to the MAN location, the results in table 6.12 show that when NO_2 emission increases, then the disease activity score for RA patients will significantly increase by 2.67 after one lag (one day) from NO_2 emission increases, and the reason for that is because MAN location is surrounded by four major traffic highways, and in addition, MAN is very close to an industrial area which is in the centre of Kuwait.

Similarly, the third finding is about the direction of the relationship between $\log(DAS28)$ and O_3 . As for the results of NO_2 , the results show a positive significant relationship between $\log(DAS28)$ and O_3 in the JAH location. When O_3 emission increases, then the disease activity score for RA patients will significantly increase by 1.875 after two lags (two days) from O_3 emission increases.

Finally, the weather factors (Temp, RH and WS) do not show any positive significant relationship with $\log(DAS28)$ among RA patients in Kuwait.

Table 6.12: The results of estimation and verification of the vector autoregressive model: VAR estimates to predict $Y_t = (DAS28_t, SO_{2,t}, NO_{2,t}, O_{3,t})'$ based on the information of air pollutants among the air monitoring fixed stations. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$.

Dependent = DAS28	ASA t-ratio	FAH t-ratio	MAN t-ratio	JAH t-ratio
const	-0.6396	-0.185141	-5.09434	-0.3521
$\log(DAS28)$ lag 1	14.21***	14.47***	14.01***	14.42***
$\log(DAS28)$ lag 2	-1.586	-1.505	-1.927*	-1.884*
$\log(DAS28)$ lag 3	0.7488	0.9152	0.6163	0.6653

Table 6.12: The results of estimation and verification of the vector autoregressive model: VAR estimates to predict $Y_t = (\text{DAS28}_t, \text{SO}_{2,t}, \text{NO}_{2,t}, \text{O}_{3,t})'$ based on the information of air pollutants among the air monitoring fixed stations. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$.

Dependent = DAS28	ASA t-ratio	FAH t-ratio	MAN t-ratio	JAH t-ratio
log(DAS28) lag 4	0.3547	0.676	0.1871	0.3427
log(DAS28) lag 5	-1.918*	-1.706*	-1.925*	-1.673*
log(DAS28) lag 6	5.329***	5.692***	5.549***	5.783***
log(DAS28) lag 7	16.53***	16.82***	16.82***	16.77***
SO_2 lag 1	-0.9181	-1.310	-0.1984	0.3453
SO_2 lag 2	-0.4475	1.905*	-0.9284	2.118**
SO_2 lag 3	1.224	-1.343	0.4937	-2.283**
SO_2 lag 4	-0.4453	0.4909	1.421	2.78***
SO_2 lag 5	0.5426	1.499	0.5997	-0.7575
SO_2 lag 6	-0.004409	-0.2892	1.438	0.4997
SO_2 lag 7	-0.6928	-0.3040	0.7391	0.8444
NO_2 lag 1	1.233	0.9038	2.67***	1.132
NO_2 lag 2	-2.089**	-1.894*	-0.6797	0.8065
NO_2 lag 3	0.9671	1.049	0.8027	-0.1255
NO_2 lag 4	-0.6236	0.4131	0.4441	0.9369
NO_2 lag 5	0.2404	-1.508	0.09582	-1.704*
NO_2 lag 6	0.5421	-0.3719	0.4579	-0.7098
NO_2 lag 7	1.004	1.205	0.6297	-1.376
O_3 lag 1	-2.192**	-2.055**	-0.6607	-2.077**
O_3 lag 2	0.3603	0.6921	-0.6951	1.875*
O_3 lag 3	0.1801	-0.3352	-0.06933	-0.1311
O_3 lag 4	1.413	2.074**	0.7754	-0.2799
O_3 lag 5	-0.4190	-1.826*	-0.5828	-0.9562
O_3 lag 6	-1.205	-0.6435	-0.5605	1.319

Table 6.12: The results of estimation and verification of the vector autoregressive model: VAR estimates to predict $Y_t = (\text{DAS28}_t, \text{SO}_{2,t}, \text{NO}_{2,t}, \text{O}_{3,t})'$ based on the information of air pollutants among the air monitoring fixed stations. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$.

Dependent = DAS28	ASA t-ratio	FAH t-ratio	MAN t-ratio	JAH t-ratio
O_3 lag 7	-1.173	-0.2936	0.3191	-0.8976
Temp	-1.223	-0.6413	0.576	-1.819*
Temp lag 1	-0.3555	0.6319	-0.8024	0.07723
Temp lag 2	1.515	-0.3139	0.5781	0.9921
Temp lag 3	-0.7100	0.7163	-1.521	-0.8275
Temp lag 4	0.2598	-1.710*	1.145	0.503
Temp lag 5	0.2101	1.424	-0.2844	-0.01190
Temp lag 6	-0.1633	-0.05875	1.03	-0.5192
Temp lag 7	0.2306	-0.1634	-0.6590	1.253
RH	-0.8138	-1.776*	0.09794	-1.588
RH lag 1	0.101	1.495	-1.007	-0.2386
RH lag 2	0.8028	-1.867*	1.589	-0.06524
RH lag 3	0.05713	1.948*	-1.718*	0.1011
RH lag 4	0.4094	-1.080	2.393**	1.393
RH lag 5	-0.2113	1.244	-2.341**	-0.4872
RH lag 6	-0.8290	-1.310	-0.2942	-0.4780
RH lag 7	0.5406	0.8546	1.615	0.3491
WS	-0.07126	-0.06144	-1.254	-0.4758
WS lag 1	0.7804	-0.3803	1.137	-0.2261
WS lag 2	-0.7738	-0.9771	0.3537	-0.02841
WS lag 3	1.167	0.3892	-0.5036	-0.5103
WS lag 4	-2.698***	-1.034	-0.4940	0.6195
WS lag 5	0.6093	-0.8763	-1.082	-0.7711
WS lag 6	0.5465	0.7003	1.105	-2.033**

Table 6.12: The results of estimation and verification of the vector autoregressive model: VAR estimates to predict $Y_t = (DAS28_t, SO_{2,t}, NO_{2,t}, O_{3,t})'$ based on the information of air pollutants among the air monitoring fixed stations. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$.

Dependent = DAS28	ASA t-ratio	FAH t-ratio	MAN t-ratio	JAH t-ratio
WS lag 7	0.6948	-0.1648	0.4803	-1.424

6.16 Vector Error Correction Model (VECM) Analysis

According to Engle-Granger (1987), if the Granger causality test is conducted at $I(1)$ using the VAR framework, it will be misrepresentative in the presence of cointegration; thus, adding another variable to the VAR method, like the error-correction term, will aid in the exploration of the long-run association. The negative coefficient of the one lagged error-correction component of the long-run effects can be used to establish the direction of causation between the fundamental variables. The VECM model treats each variable individually as endogenous, therefore the number of variables matches the number of equations in the model (Hondroyannis et al., 2002).

Each dependent variable in the VECM technique is a function of its own error-correction term, lags, a random variable and lags of explanatory variables. As a result, the VECM aids in the identification of causation among cointegrated variables as well as the detection of short and long-run correlations. The Granger-Causality test in VECM outline between $\log(DAS28)$, VECM tests the relationship between $\log(DAS28)$ and the emission of the air pollutant variables (SO_2 , NO_2 and O_3). Whereas, the other climatological variables are considered as exogenous variables. As it has previously been mentioned in Chapter 5, the VECM model is written as follows:

$$\begin{aligned}
& \begin{bmatrix} \Delta \log(DAS28_t) \\ \Delta SO_{2,t} \\ \Delta NO_{2,t} \\ \Delta O_{3,t} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix} + \begin{bmatrix} \beta_{11,1} & \beta_{12,1} & \beta_{13,1} & \beta_{14,1} \\ \beta_{21,1} & \beta_{22,1} & \beta_{23,1} & \beta_{24,1} \\ \beta_{31,1} & \beta_{32,1} & \beta_{33,1} & \beta_{34,1} \\ \beta_{41,1} & \beta_{42,1} & \beta_{43,1} & \beta_{44,1} \end{bmatrix} \begin{bmatrix} \Delta \log(DAS28_{t-1}) \\ \Delta SO_{2,t-1} \\ \Delta NO_{2,t-1} \\ \Delta O_{3,t-1} \end{bmatrix} \\
& + \dots + \begin{bmatrix} \beta_{11,k} & \beta_{12,k} & \beta_{13,k} & \beta_{14,k} \\ \beta_{21,k} & \beta_{22,k} & \beta_{23,k} & \beta_{24,k} \\ \beta_{31,k} & \beta_{32,k} & \beta_{33,k} & \beta_{34,k} \\ \beta_{41,k} & \beta_{42,k} & \beta_{43,k} & \beta_{44,k} \end{bmatrix} \begin{bmatrix} \Delta \log(DAS28_{t-k}) \\ \Delta SO_{2,t-k} \\ \Delta NO_{2,t-k} \\ \Delta O_{3,t-k} \end{bmatrix} \\
& + \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \\ \varepsilon_{4t} \end{bmatrix}.
\end{aligned} \tag{6.8}$$

It is assumed that residual terms ε_{it} are distributed independently and normally, with constant variance and zero mean. The VECM distinguishes between three categories of causality: strong, weak and long-run causal linkages (Zambrano-Monserrate et al., 2016). The following results represent the VECM with unrestricted constant, referring to equation (5.70) from Chapter 5:

Table 6.13: VECM model with lag order equal to 7 and Maximum likelihood estimates, observations from 2013-01-08 to 2020-12-31 with cointegration rank = 3 and using an unrestricted constant. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$.

Dependent = DAS28	ASA t-stat.	FAH t-stat.	MAN t-stat.	JAH t-stat.
const	-2.502**	0.3148	-5.898***	-0.5564
$\Delta \log(DAS28)_{t-1}$	-11.66***	-12.20***	-11.39***	-11.41***
$\Delta \log(DAS28)_{t-2}$	-13.43***	-14.08***	-13.47***	-13.51***
$\Delta \log(DAS28)_{t-3}$	-13.89***	-14.48***	-14.10***	-14.17***
$\Delta \log(DAS28)_{t-4}$	-15.02***	-15.46***	-15.43***	-15.44***

Table 6.13: VECM model with lag order equal to 7 and Maximum likelihood estimates, observations from 2013-01-08 to 2020-12-31 with cointegration rank = 3 and using an unrestricted constant. Note that: * p<.05, ** p<.01, *** p<.001.

Dependent = DAS28	ASA t-stat.	FAH t-stat.	MAN t-stat.	JAH t-stat.
$\Delta \log(DAS28)_{t-5}$	-18.90***	-19.28***	-19.37***	-19.28***
$\Delta \log(DAS28)_{t-6}$	-16.84***	-16.96***	-17.10***	-16.89***
$\Delta SO_{2,t-1}$	1.009	-1.719*	-2.341**	-2.491**
$\Delta SO_{2,t-2}$	0.4775	-0.2370	-2.958***	-0.9516
$\Delta SO_{2,t-3}$	1.25	-1.220	-2.661***	-2.570**
$\Delta SO_{2,t-4}$	0.7893	-0.8624	-1.784*	-0.5689
$\Delta SO_{2,t-5}$	1.128	0.3869	-1.568	-1.224
$\Delta SO_{2,t-6}$	1.084	0.2198	-0.5927	-0.9288
$\Delta NO_{2,t-1}$	-2.192**	1.095	-2.898***	2.813***
$\Delta NO_{2,t-2}$	-3.548***	-0.3642	-3.012***	3.02***
$\Delta NO_{2,t-3}$	-2.302**	0.3928	-2.087**	2.682***
$\Delta NO_{2,t-4}$	-2.581***	0.7104	-1.578	3.384***
$\Delta NO_{2,t-5}$	-2.198**	-0.5770	-1.414	2.014**
$\Delta NO_{2,t-6}$	-1.553	-1.147	-0.9488	1.542
$\Delta O_{3,t-1}$	1.013	1.163	1.412	-0.9155
$\Delta O_{3,t-2}$	1.306	1.637	0.5927	1.007
$\Delta O_{3,t-3}$	1.413	1.153	0.4199	0.8532
$\Delta O_{3,t-4}$	2.79***	3.089***	1.137	0.5554
$\Delta O_{3,t-5}$	2.486**	1.241	0.5026	-0.5090
$\Delta O_{3,t-6}$	1.318	0.5025	-0.1958	0.996
$Temp_t$	-1.091	-0.5708	0.6927	-1.789*
$Temp_{t-1}$	-0.3795	0.6228	-0.8152	0.09788
$Temp_{t-2}$	1.539	-0.3043	0.6038	1
$Temp_{t-3}$	-0.6967	0.7229	-1.510	-0.8151
$Temp_{t-4}$	0.2787	-1.713*	1.164	0.5038

Table 6.13: VECM model with lag order equal to 7 and Maximum likelihood estimates, observations from 2013-01-08 to 2020-12-31 with cointegration rank = 3 and using an unrestricted constant. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$.

Dependent = DAS28	ASA t-stat.	FAH t-stat.	MAN t-stat.	JAH t-stat.
$Temp_{t-5}$	0.2027	1.426	-0.2637	-0.005042
$Temp_{t-6}$	-0.2069	-0.06470	1.017	-0.5123
RH_t	-0.8593	-1.921*	0.05043	-1.575
RH_{t-1}	0.06087	1.479	-1.021	-0.2264
RH_{t-2}	0.798	-1.902*	1.591	-0.06244
RH_{t-3}	0.06348	1.912*	-1.722*	0.1085
RH_{t-4}	0.4036	-1.113	2.394**	1.401
RH_{t-5}	-0.2227	1.206	-2.357**	-0.4753
RH_{t-6}	-0.8588	-1.338	-0.3054	-0.4674
WS_t	-0.07329	-0.1150	-1.207	-0.5146
WS_{t-1}	0.891	-0.3754	1.376	-0.3045
WS_{t-2}	-0.6496	-0.9479	0.5708	-0.1044
WS_{t-3}	1.287	0.4284	-0.2915	-0.5804
WS_{t-4}	-2.590***	-1.006	-0.2809	0.5552
WS_{t-5}	0.7195	-0.8554	-0.8754	-0.8398
WS_{t-6}	0.6746	0.7572	1.323	-2.102**
$ECT1$	-11.02***	-11.19***	-11.94***	-11.49***
$ECT2$	-2.101**	0.9439	2.613***	3.097***
$ECT3$	7.98***	-0.5594	8.206***	-4.549***

The results in table 6.13 show four different VECM models (i.e. each VECM model refers to a separate location [ASA, FAH, MAN and JAH]). The results in table 6.14 show the VECM estimates for the ASA location. We can easily conclude that the short run effect for O_3 with lag 4 and 5 has a positive relationship with the disease activity score for RA patients, which means, if the Ozone emission tends to increase by 0.234

(VECM coefficient estimate for O_3 with lag equal to 4) there is a 1% increase in RA DAS28 and if the Ozone emission tends to increase by 0.195 with lag equal to 5 there is a 1% increase in RA DAS28, while NO_2 exerted a negative influence on the disease activity score for RA patients, that tends to decrease by around 1% if the emission of NO_2 increases by 1.71 with 2 lags.

The speed of the change that restores equilibrium in the dynamic model is shown by the error correction term (ECT) in tables 6.14 and 6.15. The ECM coefficient indicates how rapidly variables return to equilibrium, and it should have a statistically significant coefficient with a negative sign at the 5% level of significance (Pahlavani et al., 2005). The one-lag error correction terms (ECT) are found to be statistically significant and have the predicted negative sign. This demonstrates that the variables in the model have a co-integrated association. For instance, ECT1 of -0.36 (table 6.14) reveals that 36% of the discrepancy between the actual and the predictive value of the overall relationship between the disease activity scores (DAS28) is affected by air pollutants and weather factors. This implies that convergence to equilibrium is relatively high and deviations from the long run DAS28 (dependent variable) are corrected by 36% over the following year.

However, this result does not hold in the long run, when air pollutants significantly increase disease activity scores (DAS28). On the other side, the long-term period, it indicates that the relationship between the environmental quality index SO_2 , NO_2 , O_3 and disease activity scores (DAS28) has a long-term equilibrium. It shows in the values of EC1 and EC2 (Cointegration with rank 1 and 2) which are negative and significant.

The results in table 6.15 show the VECM estimates for the JAH location. The results for JAH show different indications than the other locations (e.g., if you look at table 6.14, NO_2 has a positive effect to increase the RA disease scores, however other locations indicate that NO_2 has a negative influence on RA disease scores). Because the Al-Jahra area (JAH) is bounded by various utility industries, electricity and desalination plants, the northern oil fields, roads connecting it to the rest of Kuwait and other nations, and, a wastewater treatment plant, the results are more realistic. The results in table 6.16

show that the short run effect for NO_2 with lags 1 to 4 has a positive relationship with the disease activity score for RA patients which means, if the NO_2 emissions tend to increase by 0.0182 with lag equal 2 there is a 1% increase in RA DAS28 and if the Ozone emissions tend to increase by 0.021 with lag 4 there is a 1% increase in RA DAS28.

There were several studies done by many scholars which proved and confirmed the pollution situation in Al-Jahra with increased emissions for SO_2 , NO_2 and O_3 (Al-Fadhli et al., 2019; Al-Baroud et al., 2012; Alenezi and Al-Anezi, 2015). Furthermore, Al-Hemoud et al. (2021) confirmed that ambient NO_2 levels in Kuwait surpassed both Kuwaiti EPA requirements and WHO norms. The O_3 levels were found to be extremely low, well below local and international standards. The highest NO_2 levels were observed in the early morning and mid-afternoon, during autumn and winter, and on Saturdays (the "weekend effect"). The highest O_3 levels were reported in the early morning and mid-afternoon, during autumn and winter, and on Saturdays (the "weekend effect"). Long-term and short-term NO_2 pollution exposures were found to be linked to all-cause mortality and hospital admissions for respiratory illnesses, respectively. The inversion conditions that occur throughout the evenings in Kuwait contribute to the higher nocturnal NO_2 accumulation (Al-Hemoud et al., 2018). The two main electric-water power plants (Al-Doha and Al-Zour) may be contributing to the elevated NO_2 levels observed late at night. Kuwaiti power plants use a mix of heavy fuel oil, crude oil, natural gas and gas oil, which contributes to NO_2 levels above international regulations (Al-Fadhli et al., 2019).

Table 6.14: VECM estimates – "Air pollutants" model dependent variable: $\Delta \log(DAS28)_t$ for ASA location using an unrestricted constant. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$.

VECM ASA	coefficient	std. error	t-ratio	p-value	level
const	-2.39196	0.956	-2.502	0.0124	**
$\Delta \log(DAS28)_{t-1}$	-0.379436	0.0325473	-11.66	1.02E-30	***
$\Delta \log(DAS28)_{t-2}$	-0.403711	0.0300517	-13.43	6.05E-40	***
$\Delta \log(DAS28)_{t-3}$	-0.384838	0.0277112	-13.89	1.75E-42	***

$\Delta \log(DAS28)_{t-4}$	-0.373680	0.0248787	-15.02	3.85E-49	***
$\Delta \log(DAS28)_{t-5}$	-0.404401	0.021392	-18.90	3.32E-75	***
$\Delta \log(DAS28)_{t-6}$	-0.301519	0.0179001	-16.84	8.85E-61	***
$\Delta SO_{2,t-1}$	0.193205	0.191561	1.009	0.3133	
$\Delta SO_{2,t-2}$	0.0930696	0.194893	0.4775	0.633	
$\Delta SO_{2,t-3}$	0.237655	0.190071	1.25	0.2113	
$\Delta SO_{2,t-4}$	0.141892	0.179761	0.7893	0.43	
$\Delta SO_{2,t-5}$	0.184753	0.163814	1.128	0.2595	
$\Delta SO_{2,t-6}$	0.145779	0.134467	1.084	0.2784	
$\Delta NO_{2,t-1}$	-0.652572	0.297769	-2.192	0.0285	**
$\Delta NO_{2,t-2}$	-1.17133	0.330091	-3.548	0.0004	***
$\Delta NO_{2,t-3}$	-0.785332	0.341156	-2.302	0.0214	**
$\Delta NO_{2,t-4}$	-0.870410	0.337231	-2.581	0.0099	***
$\Delta NO_{2,t-5}$	-0.699384	0.318242	-2.198	0.0281	**
$\Delta NO_{2,t-6}$	-0.422541	0.272041	-1.553	0.1205	
$\Delta O_{3,t-1}$	0.0865264	0.085403	1.013	0.3111	
$\Delta O_{3,t-2}$	0.112582	0.0861857	1.306	0.1916	
$\Delta O_{3,t-3}$	0.122085	0.0863903	1.413	0.1577	
$\Delta O_{3,t-4}$	0.234276	0.0839791	2.79	0.0053	***
$\Delta O_{3,t-5}$	0.195243	0.0785508	2.486	0.013	**
$\Delta O_{3,t-6}$	0.0942837	0.071536	1.318	0.1876	
<i>ECT1</i>	-0.360731	0.0327376	-11.02	1.11E-27	***
<i>ECT2</i>	-0.361791	0.172159	-2.101	0.0357	**
<i>ECT3</i>	1.15157	0.144299	7.98	2.09E-15	***

Table 6.15: VECM estimates – "Air pollutants" model dependent variable: $D \log(DAS28)_t$ for JAH location using an unrestricted constant.

VECM JAH	coefficient	std. error	t-ratio	p-value	level
const	-0.442770	0.513019	-0.8631	3.88E-01	
$\Delta \log(DAS28)_{t-1}$	-0.366271	0.0323887	-11.31	4.86E-29	***
$\Delta \log(DAS28)_{t-2}$	-0.399408	0.0298043	-13.40	9.16E-40	***
$\Delta \log(DAS28)_{t-3}$	-0.385873	0.0274014	-14.08	1.34E-43	***
$\Delta \log(DAS28)_{t-4}$	-0.378186	0.0246069	-15.37	2.77E-51	***
$\Delta \log(DAS28)_{t-5}$	-0.407082	0.0211975	-19.20	2.04E-77	***
$\Delta \log(DAS28)_{t-6}$	-0.299348	0.0177942	-16.82	1.23E-60	***
$\Delta SO_{2,t-1}$	-0.315700	0.133203	-2.370	0.0179	**
$\Delta SO_{2,t-2}$	-0.114534	0.13448	-0.8517	3.95E-01	
$\Delta SO_{2,t-3}$	-0.330972	0.130512	-2.536	0.0113	**
$\Delta SO_{2,t-4}$	-0.0677819	0.122644	-0.5527	0.5805	
$\Delta SO_{2,t-5}$	-0.136022	0.110428	-1.232	0.2181	
$\Delta SO_{2,t-6}$	-0.0850799	0.0892041	-0.9538	0.3403	
$\Delta NO_{2,t-1}$	0.0142354	0.00609843	2.334	0.0197	**
$\Delta NO_{2,t-2}$	0.0181633	0.00695515	2.611	0.0091	***
$\Delta NO_{2,t-3}$	0.0167398	0.00725146	2.308	0.021	**
$\Delta NO_{2,t-4}$	0.0212342	0.00707505	3.001	0.0027	***
$\Delta NO_{2,t-5}$	0.010216	0.00646597	1.58	0.1142	
$\Delta NO_{2,t-6}$	0.00496561	0.0051761	0.9593	0.3375	
$\Delta O_{3,t-1}$	-0.0118603	0.0137424	-0.8630	0.3882	
$\Delta O_{3,t-2}$	0.0148023	0.0139532	1.061	0.2888	
$\Delta O_{3,t-3}$	0.0133771	0.0138945	0.9628	0.3358	
$\Delta O_{3,t-4}$	0.00873131	0.0134905	0.6472	0.5175	
$\Delta O_{3,t-5}$	-0.00495398	0.0129027	-0.3839	0.701	
$\Delta O_{3,t-6}$	0.0125141	0.0120629	1.037	0.2996	

<i>ECT1</i>	-0.373654	0.0325914	-11.46	8.75E-30	***
<i>ECT2</i>	0.352718	0.119477	2.952	3.20E-03	***
<i>ECT3</i>	-0.00882579	0.00284637	-3.101	0.0019	***

The following results in table 6.16 show the VECM with unrestricted constant and trend (referring to equation (5.72) from Chapter 5):

Table 6.16: VECM model with lag order equal 7 and Maximum likelihood estimates, observations from 2013-01-08 to 2020-12-31 with Cointegration rank = 3 using an unrestricted constant and trend. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$.

Dependent = DAS28	ASA t-ratio	FAH t-ratio	MAN t-ratio	JAH t-ratio
const	-2.748***	0.7465	-5.033***	-0.7885
$\Delta \log(DAS28)_{t-1}$	-11.76***	-12.43***	-11.48***	-14.21***
$\Delta \log(DAS28)_{t-2}$	-13.51***	-14.29***	-13.61***	-16.05***
$\Delta \log(DAS28)_{t-3}$	-13.98***	-14.66***	-14.21***	-16.29***
$\Delta \log(DAS28)_{t-4}$	-15.09***	-15.58***	-15.55***	-17.13***
$\Delta \log(DAS28)_{t-5}$	-18.96***	-19.38***	-19.47***	-20.53***
$\Delta \log(DAS28)_{t-6}$	-16.86***	-17.00***	-17.15***	-17.51***
$\Delta SO_{2,t-1}$	1.179	-1.644	-2.130**	-4.566***
$\Delta SO_{2,t-2}$	0.6619	-0.2003	-2.805***	-2.524**
$\Delta SO_{2,t-3}$	1.41	-1.166	-2.539**	-3.957***
$\Delta SO_{2,t-4}$	0.9302	-0.8097	-1.706*	-1.649*
$\Delta SO_{2,t-5}$	1.247	0.4778	-1.456	-2.100**
$\Delta SO_{2,t-6}$	1.198	0.312	-0.5677	-1.540
$\Delta NO_{2,t-1}$	-2.366**	1.5	-2.919***	3.087***
$\Delta NO_{2,t-2}$	-3.667***	-0.04629	-2.997***	3.257***
$\Delta NO_{2,t-3}$	-2.350**	0.7018	-2.037**	2.915***
$\Delta NO_{2,t-4}$	-2.581***	1.008	-1.457	3.479***
$\Delta NO_{2,t-5}$	-2.127**	-0.3771	-1.233	1.988**

Table 6.16: VECM model with lag order equal 7 and Maximum likelihood estimates, observations from 2013-01-08 to 2020-12-31 with Cointegration rank = 3 using an unrestricted constant and trend. Note that: * p<.05, ** p<.01, *** p<.001.

Dependent = DAS28	ASA t-ratio	FAH t-ratio	MAN t-ratio	JAH t-ratio
$\Delta NO_{2,t-6}$	-1.490	-1.075	-0.6091	1.26
$\Delta O_{3,t-1}$	0.8597	1.553	1.462	-0.2986
$\Delta O_{3,t-2}$	1.179	1.966**	0.6203	1.526
$\Delta O_{3,t-3}$	1.291	1.436	0.3885	1.365
$\Delta O_{3,t-4}$	2.665***	3.305***	1.119	0.991
$\Delta O_{3,t-5}$	2.35**	1.398	0.4537	-0.1164
$\Delta O_{3,t-6}$	1.174	0.5221	-0.3210	1.226
$Temp_t$	-1.093	-0.5549	0.7461	-1.741*
$Temp_{t-1}$	-0.3800	0.6446	-0.8849	0.07058
$Temp_{t-2}$	1.557	-0.2884	0.5894	1.139
$Temp_{t-3}$	-0.7163	0.6866	-1.523	-0.8564
$Temp_{t-4}$	0.3106	-1.700*	1.132	0.6216
$Temp_{t-5}$	0.1355	1.469	-0.2960	-0.08086
$Temp_{t-6}$	-0.03828	-0.3503	0.8263	0.5223
RH_t	-0.8673	-1.879*	0.1158	-1.738*
RH_{t-1}	0.04071	1.498	-0.9441	-0.2374
RH_{t-2}	0.808	-1.892*	1.55	0.02475
RH_{t-3}	0.08635	1.932*	-1.589	0.07907
RH_{t-4}	0.4109	-1.045	2.447**	1.416
RH_{t-5}	-0.2689	1.099	-2.278**	-0.3919
RH_{t-6}	-0.7272	-1.100	0.6102	-0.1214
WS_t	0.02792	0.01191	-1.158	-0.5807
WS_{t-1}	0.9842	-0.3093	1.368	-0.4774
WS_{t-2}	-0.5550	-0.8493	0.6386	-0.2132
WS_{t-3}	1.434	0.4828	-0.2231	-0.7583

Table 6.16: VECM model with lag order equal 7 and Maximum likelihood estimates, observations from 2013-01-08 to 2020-12-31 with Cointegration rank = 3 using an unrestricted constant and trend. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$.

Dependent = DAS28	ASA t-ratio	FAH t-ratio	MAN t-ratio	JAH t-ratio
WS_{t-4}	-2.505**	-0.9155	-0.2296	0.4599
WS_{t-5}	0.8163	-0.7751	-0.8560	-0.9825
WS_{t-6}	1.054	0.8261	1.604	-2.518**
$ECT1$	-10.94***	-11.04***	-11.92***	-10.31***
$ECT2$	-2.324**	0.8835	2.227**	6.26***
$ECT3$	8.678***	-1.123	8.146***	-3.253***

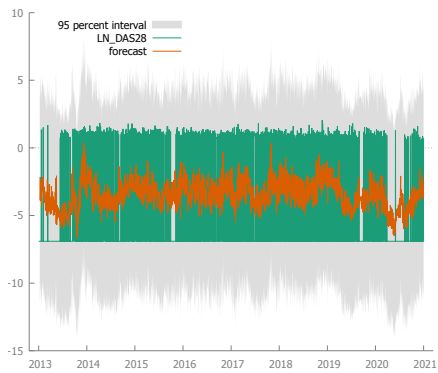
The results in table 6.16 show the same significant effect that was described from table 6.13 between RA disease activity score and air pollution. That means, when we use the unrestricted constant and trend, that will never explain the trend effect between air pollution and DAS28.

Table 6.17 shows comparisons between time series models using the information of R-Square, MAPE, RMSE, MAE, MPE and Theil's U2. It is very clear that the best model that explains and captures the most fitted values was VECM with R-square varying between 0.43 to 0.45, and with mean predictive error varying between 2.87 to 2.9 with lowest error among the other time series models in prediction performance. Also, if we look at figures 6.19 to 6.22, we can see that the best model with the best prediction performance (the time series model that captures most of the green lines) was VECM (e.g. the green line presents the actual values of DAS28, the red line presents the model forecast, and we can see that the models with the weakest performance in prediction were OLS and GARCH models because the model predictions could not capture most of the actual values). Therefore, the results revealed that the values of the VECM model of cointegration for long- and short-term RA disease activity predictions were lower than the values predicted by the time series models. It meant that the VECM model of cointegration predicting performance was sufficient, and that the model did not need

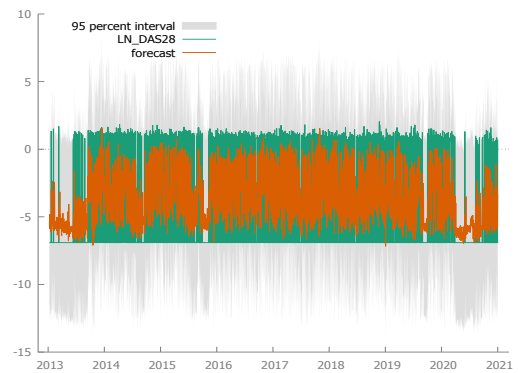
to be revised. These statistics indicated that the VECM model's predicting ability is superior compared to other time series models, and this conclusion was supported by the data (Khin et al., 2015).

Table 6.17: Comparison of models based on several selection criteria (R-Square, MAPE, RMSE, MAE, MPE and Theil's U2).

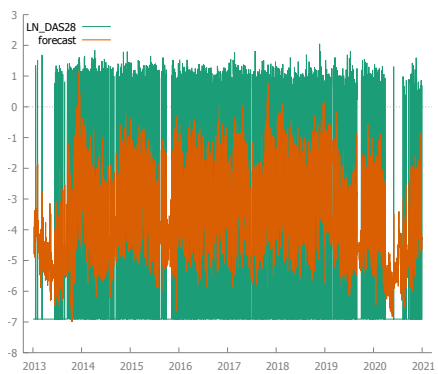
Station	Model	R-Square	MAPE	RMSE	MAE	MPE	Theil's U2
ASA	OLS	0.066	307.660	3.783	3.627	159.880	0.688
	ARIMA	0.152	271.590	3.606	3.296	152.300	0.752
	GARCH		307.920	3.822	3.552	184.180	0.846
	VAR	0.250	221.640	3.390	2.911	149.230	0.367
	VECM - Constant	0.436	215.810	3.397	2.877	151.100	0.397
	VECM - Trend	0.435	215.610	3.398	2.876	151.160	0.400
FAH	OLS	0.048	336.290	3.820	3.697	137.060	0.388
	ARIMA	0.145	292.220	3.619	3.318	132.910	0.529
	GARCH		354.170	3.883	3.615	146.380	0.477
	VAR	0.249	226.940	3.392	2.915	143.080	0.310
	VECM - Constant	0.437	223.470	3.394	2.895	143.660	0.319
	VECM - Trend	0.436	223.980	3.395	2.888	142.710	0.318
MAN	OLS	0.064	322.730	3.789	3.638	149.150	0.642
	ARIMA	0.149	283.490	3.611	3.306	143.870	0.744
	GARCH		341.080	3.842	3.561	160.760	0.565
	VAR	0.254	225.480	3.382	2.899	146.830	0.434
	VECM - Constant	0.439	221.660	3.386	2.875	147.600	0.448
	VECM - Trend	0.440	222.920	3.385	2.879	146.890	0.444
JAH	OLS	0.052	333.980	3.813	3.682	139.930	0.636
	ARIMA	0.145	287.390	3.620	3.319	139.730	0.769
	GARCH		360.050	3.879	3.639	145.140	0.548
	VAR	0.253	241.990	3.385	2.903	128.240	0.466
	VECM - Constant	0.439	240.010	3.387	2.895	128.340	0.473
	VECM - Trend	0.435	236.150	3.400	2.842	125.170	0.471



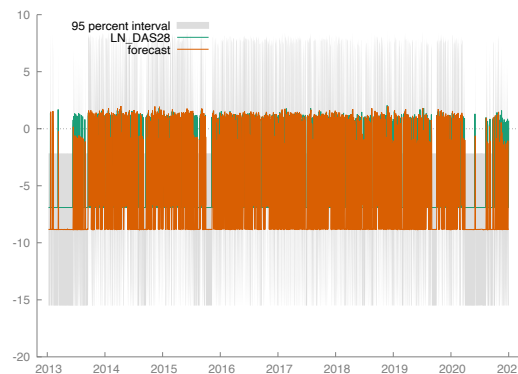
(a) OLS



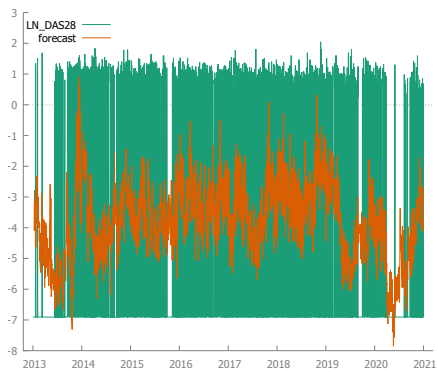
(d) VAR



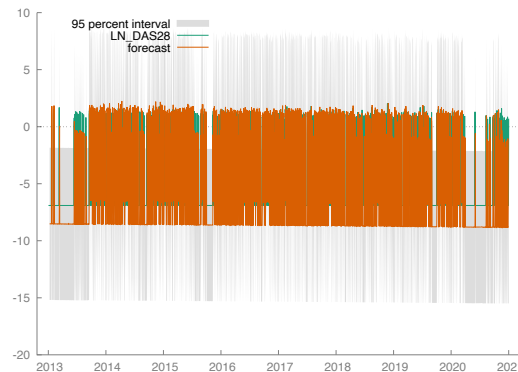
(b) ARIMA(1,0,1)



(e) VECM - Constant



(c) GARCH

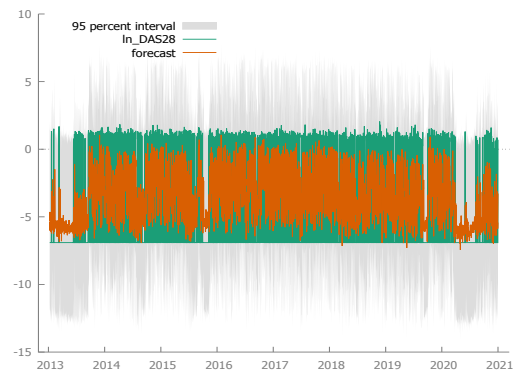


(f) VECM - Constant+Trend

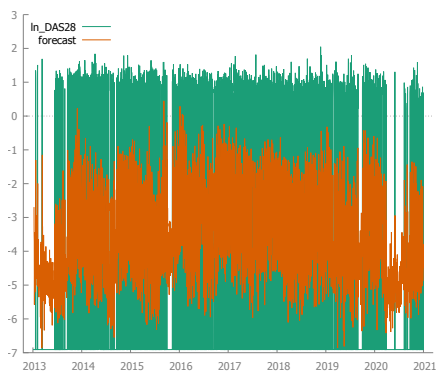
Figure 6.19: Forecast performance for the study time series models for ASA station.



(a) OLS



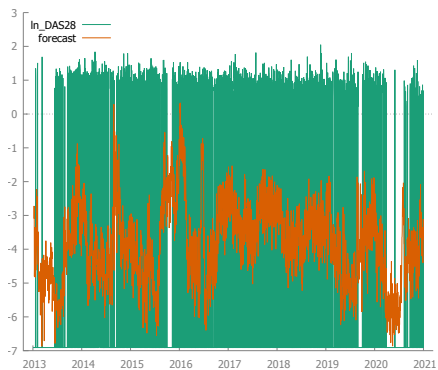
(d) VAR



(b) ARIMA(1,0,1)



(e) VECM - Constant

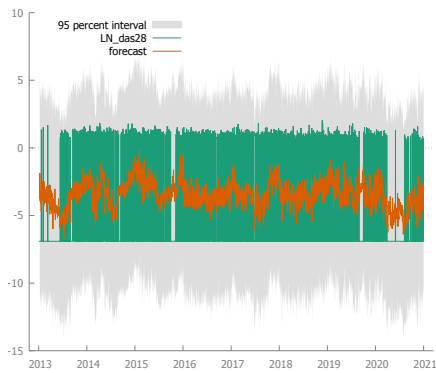


(c) GARCH

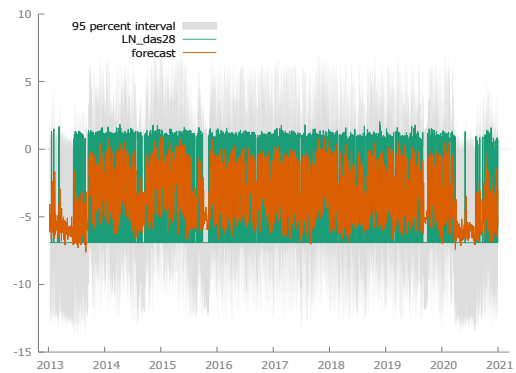


(f) VECM - Constant+Trend

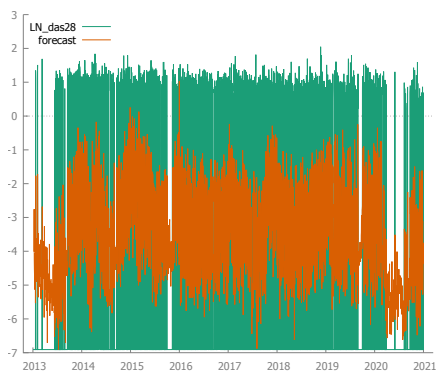
Figure 6.20: Forecast performance for the study time series models for FAH station.



(a) OLS



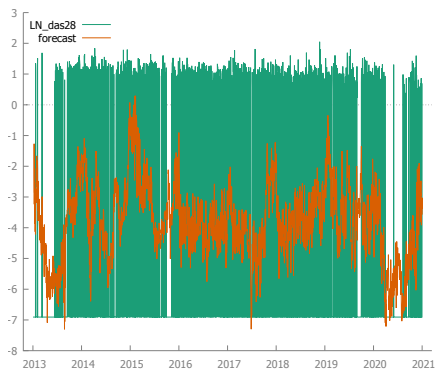
(d) VAR



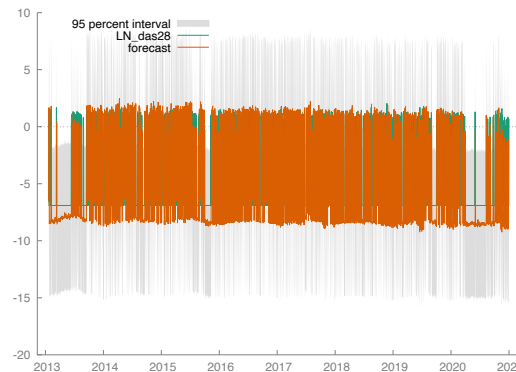
(b) ARIMA(1,0,1)



(e) VECM - Constant

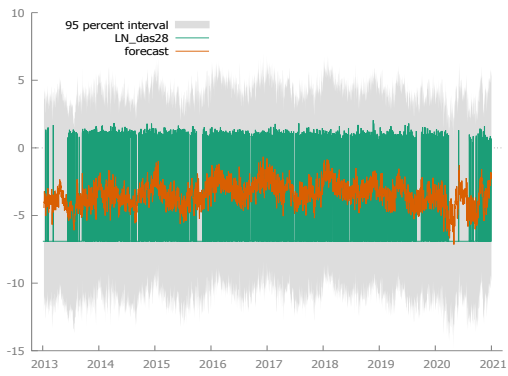


(c) GARCH

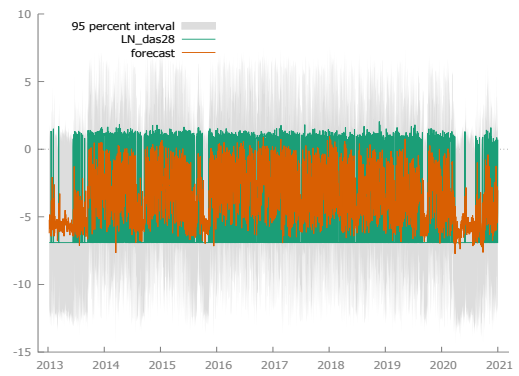


(f) VECM - Constant+Trend

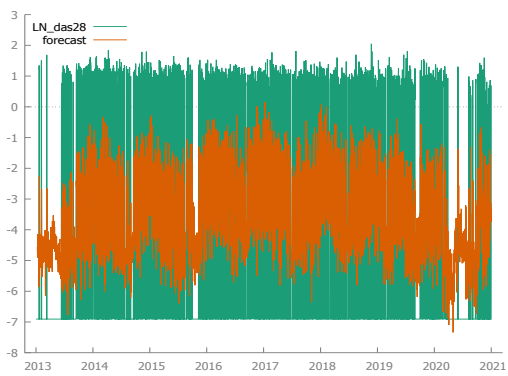
Figure 6.21: Forecast performance for the study time series models for MAN station.



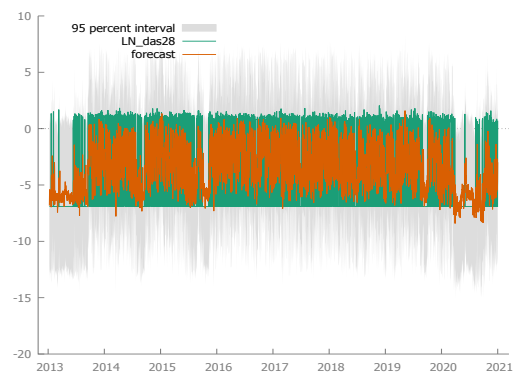
(a) OLS



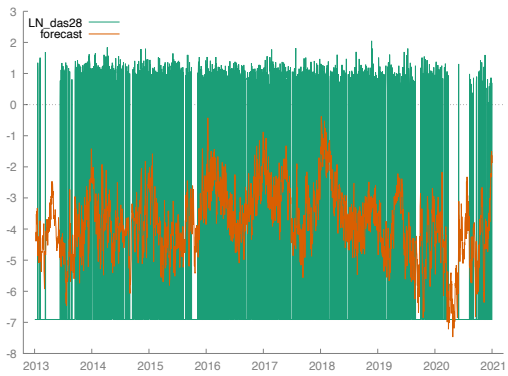
(d) VAR



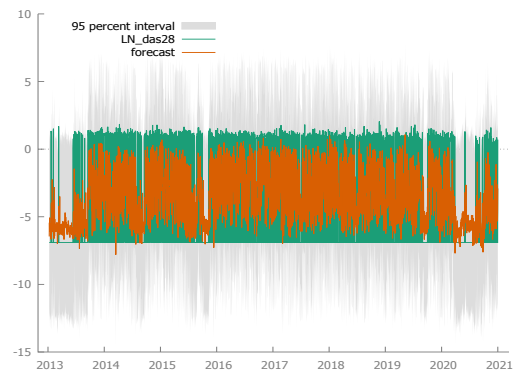
(b) ARIMA(1,0,1)



(e) VECM - Constant



(c) GARCH



(f) VECM - Constant+Trend

Figure 6.22: Forecast performance for the study time series models for JAH station

6.17 The Results of the Impulse Response Analysis

VECM Granger causality, on the other hand, is unable to provide a meaningful estimate of the strength of the causal association between variables beyond the sample period chosen. Furthermore, Granger causality only considers the direction of a causal association, rather than the sign of the association. VECM order is thought to have little effect on the Cholesky impulse response function (IRF) (Enders, 2008; Johansen, 1991). This allows the IRF to evaluate if a shock has a positive or negative long-term or short-term effect on the future and current values of all endogenous variables. The magnitude of the relevant effect is not provided by the IRF (Lau et al., 2018).

The Impulse Response Functions (IRFs) based on back-transforming the VECM model to its level VAR for each site are shown in Figures 6.23 to 6.26. The sensitivity of the dependent variables in a VAR to shocks from every variable is traced out using impulse response analysis (Brooks et al., 2008). It also depicts the effects of shocks on the variables' adjustment paths. It displays the magnitude of the shock's impact as well as the rate at which the shock dissipates, taking into account interdependencies and demonstrating how each variable reacts dynamically to shocks. The following order was used: $\log(DAS28)$, SO_2 , NO_2 and O_3 .

Figures 6.23 to 6.26 show the results of DAS28's response to its own shocks, in the first row of plots. It displays the impulse response function of DAS28's progress over a 20-day time horizon, showing the dynamic response of DAS28 to standard deviation shocks in one period to the exposure of air pollutants' emissions, as well as the persistence and direction of the response to each of its own shocks. These findings revealed that in the short run, the response of patients with RA disease activity scores to a one standard deviation of pollution in its past values was significantly positive (fundamentally from period one to around the 1st and 2nd horizons from the response from SO_2 to DAS28 in ASA) before oscillating around negative values (say the 3rd horizon to the 6th horizon), and in the long run, it is moving smoothly around zero, and results for the other locations showed the same conclusion to explain the shock response of DAS28 from SO_2 .

However, the response of DAS28 shock from NO_2 in the ASA location, showed that the response of patients with RA disease activity scores to a one standard deviation shock of the pollution in its prior values, in the short run, was significantly positive (fundamentally from period one to around the 1st and 2nd horizon from the response from NO_2 to DAS28 in ASA) prior to an oscillation around negative values (say the 5th horizon). Moreover, shocks on NO_2 emission values increase disease activity scores among RA patients (say from the 5th day horizon to more than the 20th horizon day) and this is similar with the MAN and FAH locations.

6.18 The Stability of VECM

Finally, checking the stability of the model is an important test in time series analysis. The stability of the VAR model requires the moduli of the eigenvalues to lie within the unit circle. Otherwise, the system is not stationary. Rather it is explosive or non-convergent. Figure 6.27 confirms that all roots are less than one and no root lies outside the unit circle for each model used in this study. This result signifies and satisfies the stability condition of the model. As the VAR model is stable, it is possible to present the impulse response functions and variance decomposition in response to a one-time shock in the system.

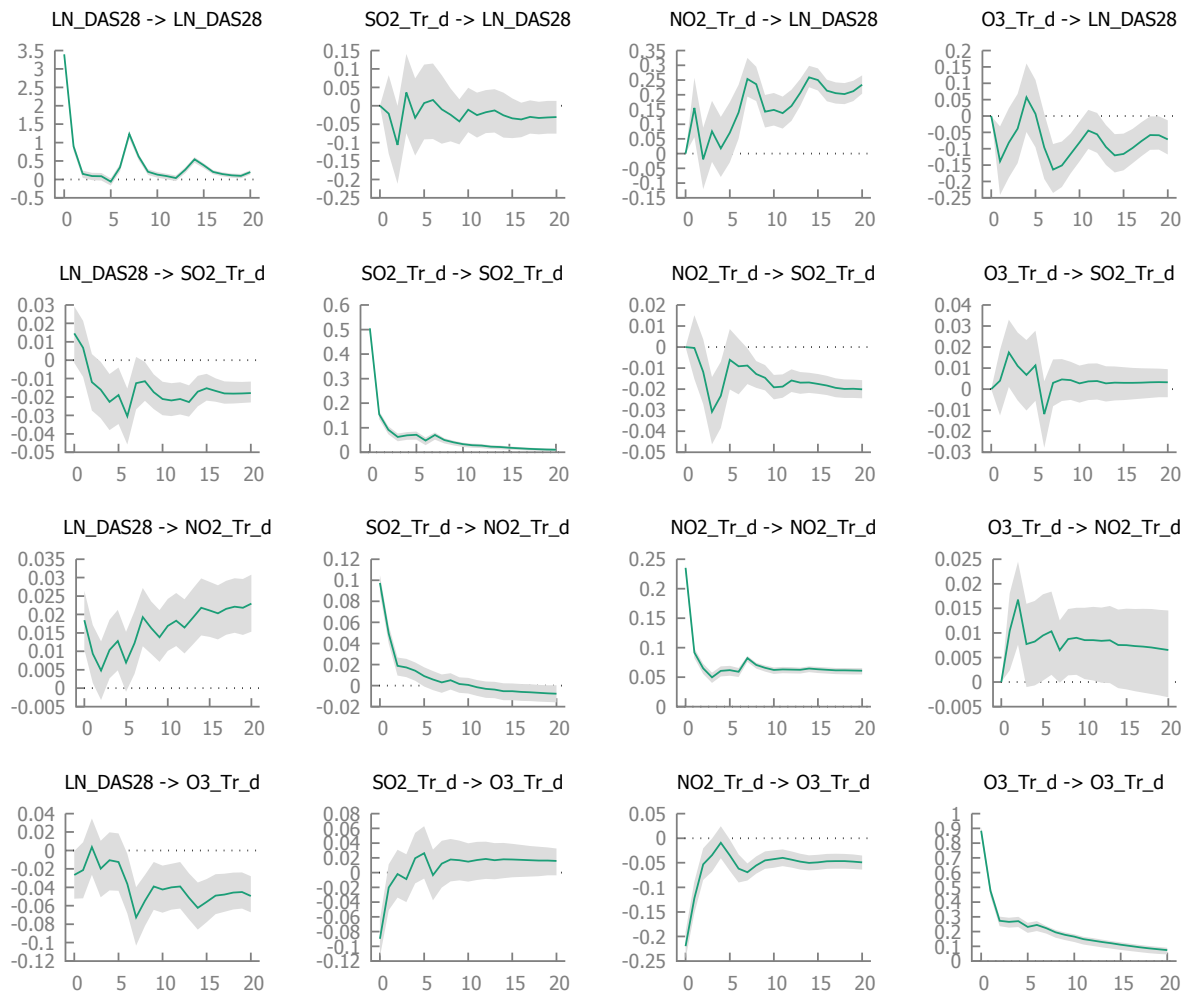


Figure 6.23: The results of the Impulse Response Functions (IRFs) based on back-transforming the VECM model to its level VAR representation for the ASA station dataset.

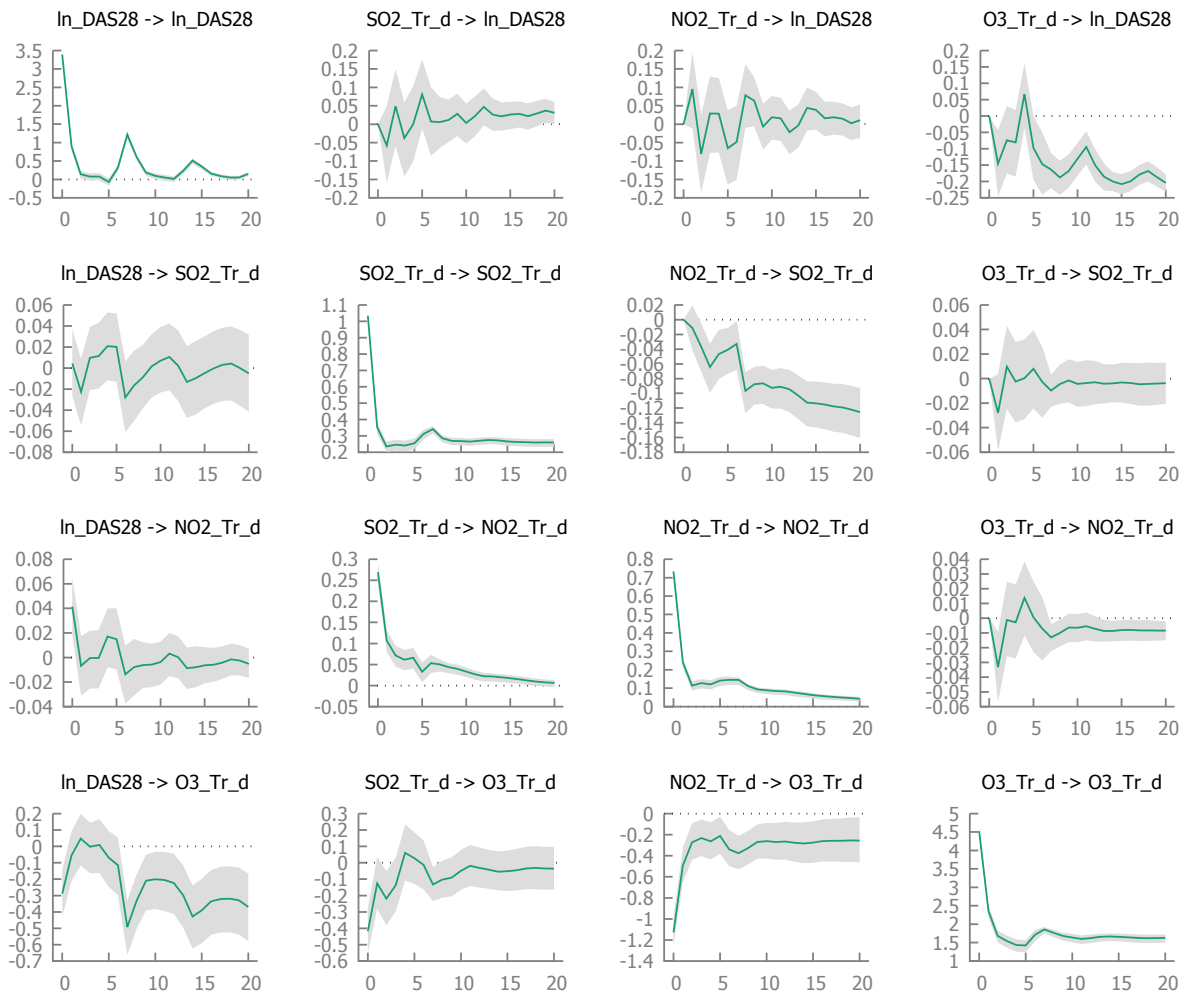


Figure 6.24: The results of the Impulse Response Functions (IRFs) based on back-transforming the VECM model to its level VAR representation for the FAH station dataset.

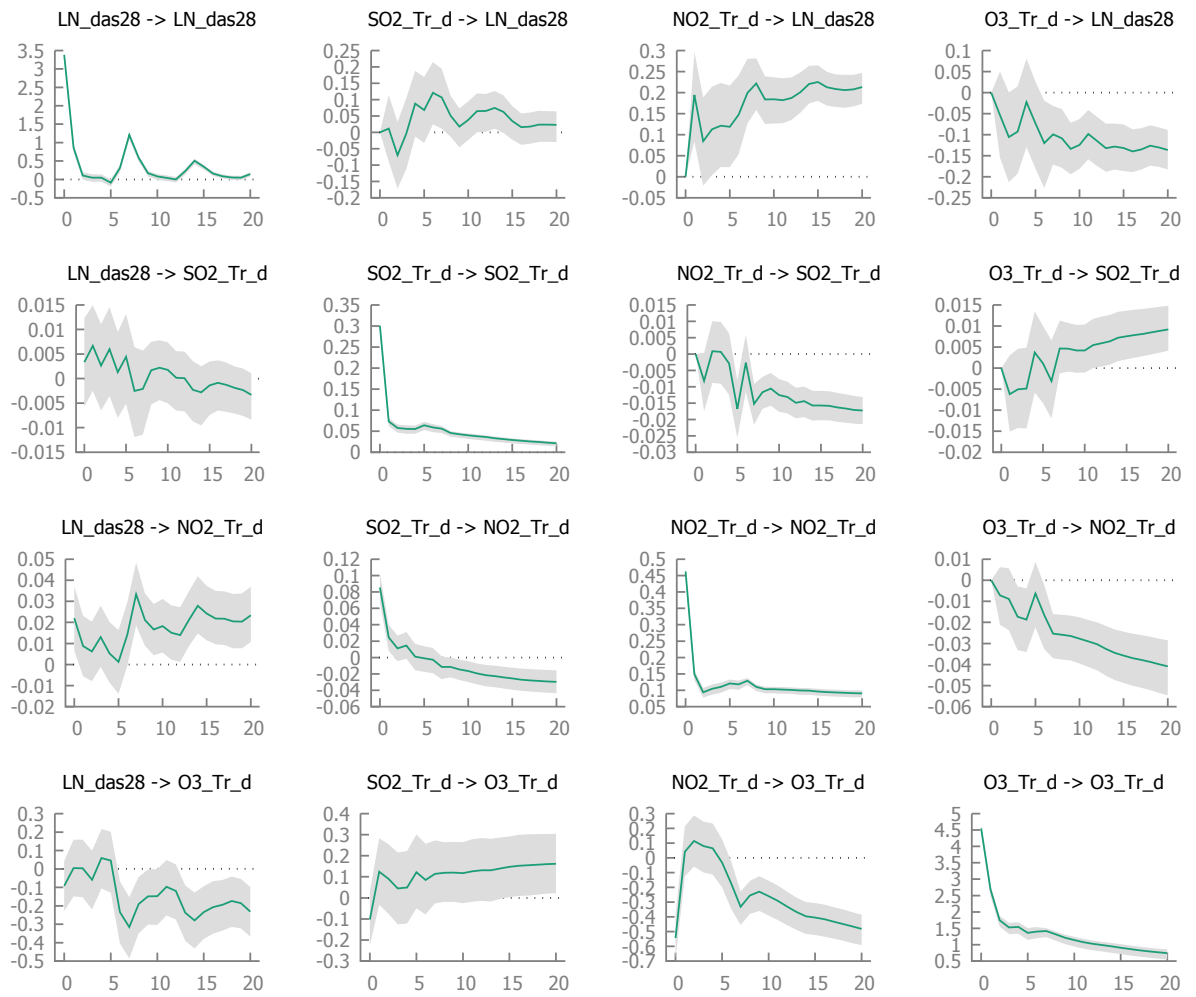


Figure 6.25: The results of the Impulse Response Functions (IRFs) based on back-transforming the VECM model to its level VAR representation for the MAN station dataset.

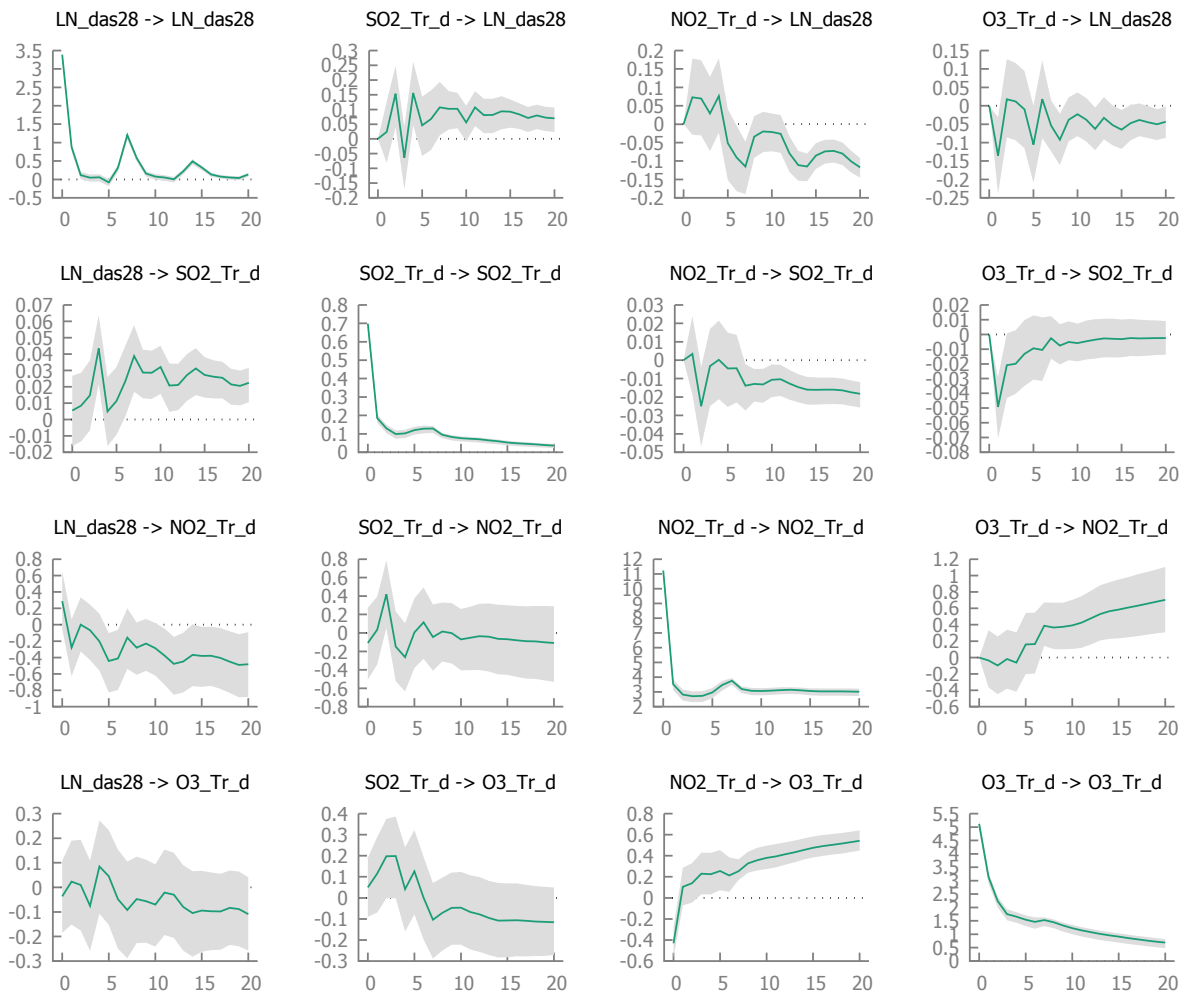


Figure 6.26: The results of the Impulse Response Functions (IRFs) based on back-transforming the VECM model to its level VAR representation for the JAH station dataset.

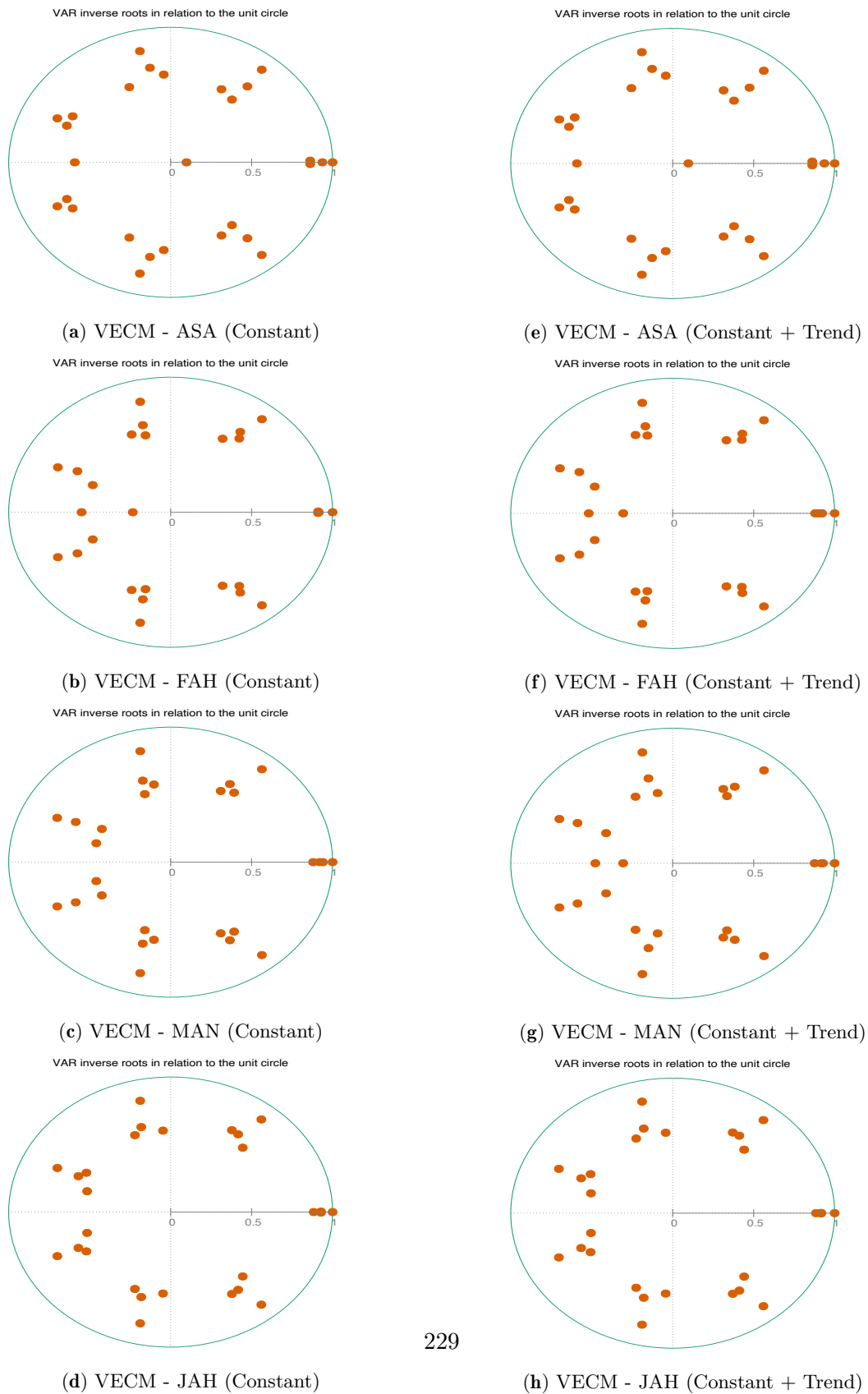


Figure 6.27: Roots of the cointegration matrix.

6.19 Discussion and Conclusion

The results of this chapter agreed with and support our previous published paper that presented in Chapter 3. In a previous work (Alsaber et al., 2020), we used numerous regression models to explain the link between RA disease activity and air pollutants, ensuring that the link remained even after the addition of RA variables that were highly significant for CDAI and DAS28. Moreover, Hart et al. (2013b) explored the effect of extended exposure to air pollution on the possibility of suffering RA in the Swedish Epidemiological Investigation. From this exploration, there was no evidence of a higher risk of RA for PM_{10} contact. Moreover, for gaseous pollutants, the overall risks for RA were mildly increased, though after controlling for the variables education and smoking, the effect was not statistically significant. The presence of a greater risk of RA incidence following increases in SO_2 and NO_2 was established in this thesis.

We also found that the most affected location is Jahra (JAH). In Jahra, the increases of air pollutants' emissions (i.e. NO_2) will positively influence DAS28 among RA patients. Based on the findings, NO_2 negatively affects DAS28 for RA patients living in the ASA and MAN locations. As we mentioned before, because Jahra is surrounded with major Al-Doha electric-water power plants that increases the emissions of NO_2 incidence throughout late nights. Moreover, these power plants usually work on a mixture of heavy fuel oil, crude oil, natural gas and gas oil that that cause NO_2 pollution. This is the reason why RA patients who belong to the JAH location have a high disease activity score of DAS28 when NO_2 increases and this agrees with Hamoda et al. (2022); Al-Fadhli et al. (2019); Al-Baroud et al. (2012).

In addition, the VECM with unrestricted constant and trend confirmed that Ozone (O_3) was detected as a risk factor for increasing the DAS28 score among the RA patients who are living in ASA and FAH. Those two locations are located in the south of Kuwait (see figure 6.1). The sources of pollution around Al-Fahaheel (FAH) and Ali Abdullah Al-Salem (ASA) locations are from the southeast (Mina Al-Ahmadi, storage tanks and refineries, 2 kilometres away; Shuaiba refinery, 4.6 kilometres away; and Mina Abdullah

refinery, 8 kilometres away), from the south (Shuaiba Industrial Area, factories, 8.8 kilometres away), and from the west (Magwaa and Burgan fields 8 kilometres away) (Hamoda et al., 2022). Those mentioned sources are the main cause of increasing the emission of O_3 for the ASA and FAH locations. Ozone has long been regarded as a key pollutant affecting air and environmental quality as a vital indicator substance of photochemical smog (Steinfeld, 1998; Seinfeld, 1989). Furthermore, the World Health Organisation (WHO) classifies six air pollutants as detrimental to human health: ground-level ozone (O_3), sulphur dioxide (SO_2), carbon monoxide (CO), lead, nitrogen dioxide, and suspended particle matter (SPM), which is commonly found in smoke and dust (Organization et al., 1997).

The one lag error correction terms (ECT) are found to have the expected negative sign and are highly statistically significant. This confirms the existence of co-integrated relationships among the variables in the model. So this indicates that the variables have a contrasting impact on DAS28 in the short run compared to the impact in the long run. O_3 and NO_2 which both exhibited a negative influence on DAS28 in the long run, exert a positive effect in the short run, especially in ASA and MAN.

So, conclusively, the VECM confirms that for long and short run equilibrium there are causal effects from O_3 and NO_2 toward the disease activity score for RA patients who are living in the residential areas surrounded by the sources of pollutants.

This study has several limitations. Firstly, there were very few published articles discussing the relation between air pollution and rheumatology disease using the disease activity scores. Furthermore, we acknowledge that one of our study's major limitations is that we did not track particle matter, which has been linked to a number of well-documented health hazards, and that was because of the missing values, with more than 60% of the total daily records missing.

Secondly, to link between RA patients' records and the information of air pollutants according to patients' living locations, we faced some problems to make that link because some patients have several living addresses, and because of that, we have dropped them out from the study.

Lastly, we could not use mobile labs to measure the daily observation of air pollutants in some residential locations that are close to the main hospitals in Kuwait, however the provided air pollution data were collected from air observation fixed stations and most of them are far away from patients' living addresses. The reason for that was because this study was not funded by any institute or governmental organisations, and all the costs and expenses were covered by the main author. Moreover, other factors, such as wind speed, humidity, and wind direction, may need to be modified in the multivariate time series using VECM. However, we did not adjust them in our study because their data was not accessible per study location or was not genuine.

Based on that, some policies and recommendations that can be given to the Kuwait government. First, there are a variety of policy solutions that can aid in the reduction of emissions. One of them is the imposition of pollution charges. Another strategy to help reduce air pollution levels is to enhance the role of renewable and clean energy, for instance, nuclear energy, consumption and energy efficiency. Suggestions for future research should employ time series techniques founded on machine learning or artificial intelligence with deep learning processes, such as the deep repeated neural network (DRNN) model, or a hybrid deep neural network (HDNN) framework which is one of the most successful to predict air pollution (Bhanja and Das, 2021). The Artificial Neural Networks (ANNs) are machine learning approaches and their basic idea is about constructing a model for mimicking the intelligence of human brain into a machine. A spatial method could be used to estimate the distance between a living area or a workplace and a pollution source. Also, crucial variables controlling the relationship between air pollution and disease activity include economic variables or social characteristics.

In this chapter, the VECM was employed to investigate the effects of air ambient pollutant (SO_2 , NO_2 , O_3 , PM_{10} and CO) emissions on a RA disease activity score (DAS28) in Kuwait over the period of 2013-2020. The empirical results show that there is a long-term cointegration relationship between SO_2 , NO_2 , O_3 and DAS28.

According to the Granger causality test and the VECM, the emissions of nitrogen dioxide (NO_2) and ozone (O_3) have a positive short-term effect on the rheumatoid

activity score (DAS28) among RA patients in Kuwait. Impulse response test results show that for some locations in Kuwait there is a short-term positive causal relationship between emissions of NO_2 and DAS28, due to sources of pollution surrounding the location. While emissions increased in NO_2 and O_3 , they increase the disease activity index (DAS28) in patients with RA from Kuwait.

Chapter 7

Multivariate Time Series: The Impact of Air Pollution on COVID-19 Daily Cases in Kuwait using the VECM Approach

In this chapter, we explore the association between air pollutant concentration rates for O_3 , SO_2 , NO_2 , CO and PM_{10} with daily COVID-19 admitted cases in Kuwait during the period 10-03-2020 to 31-12-2020 using the cointegration test and the vector error correction model (VECM).

We used a multivariate framework called the Vector Error Correction Model to create 30-days-ahead forecasts using a leading indicator, the local COVID-19 infection incidence, as well as the rising or decreasing level of daily concentrations of air pollutants (O_3 , SO_2 , NO_2 , CO , and PM_{10}). This model is also used to generate 60-day scenario estimates based on various pandemic trajectories. The two time series show a steady long-run relationship, according to our findings. In comparison to a more traditional model based solely on medical data, the model exhibits a strong fit for the data and good forecasting performance. Our study proposes a novel model for precise short-term

forecasts and practical scenario-based long-term forecasts of COVID-19 daily cases in Kuwait utilising daily air pollution concentrations (O_3 , SO_2 , NO_2 , CO , and PM_{10}) to aid healthcare decision-making.

7.1 The Relationship Between Air Pollution and COVID-19 Hospitalisation

In urban areas, air pollution is one of the most serious global environmental issues. Using time series' approaches, this study looked into the validity of the relationship between air pollution and COVID-19 hospitalisation. This time series research was carried out in the state of Kuwait. It used stationarity testing, cointegration testing, Granger causality and stability tests, and finally building the multivariate time series using the Vector Error Correction Model (VECM) technique. The findings reveal that the concentration rate of air pollutants (O_3 , SO_2 , NO_2 , CO , and PM_{10}) has an effect on COVID-19 admitted cases via Granger-causality. The Granger causation test shows that the concentration rate of some air pollutants (O_3 and SO_2) influences and predicts the COVID-19 admitted cases. The findings suggest that Ozone (O_3) and sulphur dioxide (SO_2) induce an increase in COVID-19 admitted cases in the short term. The evidence of a positive long-run association between COVID-19 admitted cases and environmental air pollution might be shown in the cointegration test and the VECM. There is a confirmation that the usage of air pollutants (O_3 , SO_2 , NO_2 , CO , and PM_{10}) has a significant impact on COVID-19 admitted cases' prediction.

Healthcare systems must have sufficient resources to meet demand from COVID-19 cases during the epidemic. One of the most essential planning measures is to examine the association between the daily cases of COVID-19 patients with the concentrations of five major air pollutants: Ozone (O_3), sulphur dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), and particulate matter (PM_{10}). Only a few articles explore the potential utility of local COVID-19 infection incidence data in developing a forecasting model for the COVID-19 hospital census using multivariate time series models (Nguyen

et al., 2021). Nguyen et al. (2021) used a multivariate framework called the Vector Error Correction Model to create 30-days-ahead forecasts using a leading indicator, the local COVID-19 infection incidence, as well as the rising or decreasing level of daily concentrations of air pollutants (O_3 , SO_2 , NO_2 , CO , and PM_{10}). This model is also used to generate 60-day scenario estimates based on various pandemic trajectories. In comparison to a more traditional model based solely on medical data, the model exhibits a better fit to the data and good forecasting performance. Our study proposes a novel model for precise short-term forecasts and practical scenario-based long-term forecasts of COVID-19 daily cases in Kuwait utilising daily air pollution concentrations (O_3 , SO_2 , NO_2 , CO , and PM_{10}) to aid healthcare decision-making.

The need for hospital administrators to have timely and precise air pollution projections in order to plan surges in hospital demand due to the epidemic spurred our effort. When hospitals surpass their historical capacity, adequate preparation can help minimise or mitigate demands on hospital resources to deal with COVID-19. As a result, a model that predicts the number of COVID-19 positive patients who will be admitted to a hospital or health system in the short and long term is critical. This COVID-19 hospital census is vital for making decisions that involve a lot of forethought, such as hiring more people, building physical beds and rooms, and purchasing critical equipment (for instance, personal protective equipment and ventilators).

Using univariate time series models such as Seasonal Autoregressive Integrated Moving Average (SARIMA), Autoregressive Integrated Moving Average (ARIMA), and exponential smoothing, past research has shown the utility of forecasting hospital demands (e.g., hospital admissions, intensive care unit census, and overall hospital census) (Earnest et al., 2005; Jones et al., 2008; Capan et al., 2016; Nguyen et al., 2021; Konaras-inghe, 2020; Yonar et al., 2020; Tyagi et al., 2020; Roy et al., 2021).

In this chapter, we seek to acquire further evidence in order to establish a link between air pollution concentrations and daily COVID-19 admitted cases in Kuwait. Our study's essential contribution and innovation are as follows:

- To our knowledge, the majority of the existing literature focuses on examining the

relationship between COVID-19 admitted cases and other climatology factors like average humidity (Fareed et al., 2020) or investigating the correlation between the average daily temperature and the rate of coronavirus epidemic growth in the affected regions (Pirouz et al., 2020). However, there is a paucity of literature that examines the association between daily COVID-19 admitted cases and air pollution.

- Most notably, this study seeks to explore the dynamic causality between air pollutants (O_3 , SO_2 , NO_2 , CO , and PM_{10}) concentrations rate and the daily COVID-19 admitted cases using the panel Granger causality test based on the vector error correction model (VECM).
- The VECM was chosen for this study for the following reasons: The method can allow endogenous variables; the VECM methodology can provide alternative analysis channels to analyse causality that is disregarded by the traditional Granger causality test due to the error correction term (ECM) (Azlina et al., 2014). Meanwhile, the VECM is capable of distinguishing between short-run and long-run causality (Azlina et al., 2014).

7.2 Literature Review

7.2.1 Relationship Between Air Pollution and Human Health

Numerous studies have indicated that the major air pollutants causing adverse health effects in Saudi Arabia include O_3 , SO_2 , NO_2 , CO and PM_{10} (Al Mulla et al., 2015; Argyropoulos et al., 2016).

It has been discovered that incomplete burning of Arabian incense produces emissions of CO , PM_{10} , $PM_{2.5}$, black carbon, and polycyclic aromatic hydrocarbons (PAHs), all of which have negative health effects on the population who are exposed to these emissions (Du et al., 2018). Ischemic heart disease (IHD), chronic obstructive pulmonary disease (COPD), and lung cancer have all been linked to these air pollutants (Amoatey et al.,

2018).

7.2.2 Impact of Air Pollution as a Risk Factor to COVID-19 patients

During COVID-19, air pollution was identified as a risk factor in several Italian research studies. Among the areas of Northern Italy, a correlation with higher levels of pollutants such as PMs has a considerable impact on human health (Domingo et al., 2020; Martelletti and Martelletti, 2020). It has also been discovered that people who live in areas with high levels of air pollution are more likely to acquire chronic respiratory illnesses and are more susceptible to any infectious agent (Distante et al., 2020).

In China, air pollution has been proven to be positively associated with SARS mortality (Cui et al., 2003). Although COVID-19 risk factors are still being investigated, it is probable that environmental variables such as air pollution could substantially impact the epidemic's spread among the population. In the case of SARS-CoV-2, many studies have found a significant relationship between air pollution and the rate at which the virus spreads. Six air pollutants ($PM_{2.5}$, PM_{10} , SO_2 , CO , NO_2 , and O_3) were significantly linked to confirmed cases in 120 Chinese cities from January 23 to February 29, 2020, according to Zhu et al. (2020). The most badly afflicted regions in Europe are the same as the ones with the highest concentrations of PM_{10} and $PM_{2.5}$, according to Martelletti and Martelletti (2020). In addition, the majority of fatality cases were in areas with the highest NO_2 concentrations (Ogen, 2020). According to Bashir et al. (2020a); Sharma et al. (2020), the associations were also confirmed in California, the United States, and India.

7.2.3 The Relationship between Atmospheric Variables and Numbers of COVID-19 Cases

Finally, for other coronavirus epidemics, it is well documented in the literature how climatic circumstances can influence transmission, either promoting or reducing it. Atmospheric variables such as ambient temperature and humidity, as well as sun irradiation, have various impacts on coronavirus survival, for example, (Casanova et al., 2010; Lauc

et al., 2020). This indicates that the coronavirus spread is facilitated in dry and cold weather. Nonetheless, it is still unknown if and how the SARS CoV-2 virus spreads or is impacted by meteorological factors like other seasonal viruses. Several recent studies looked at the role of meteorological variables in COVID-19 transmission all over the world. As shown in Pani et al. (2020) studies from China (Shi et al., 2020; Liu et al., 2020; Xie and Zhu, 2020; Ma et al., 2020), Iran (Ma et al., 2020), Spain (Briz-Redón and Serrano-Aroca, 2020), the USA (Bashir et al., 2020b; Gupta et al., 2020), Indonesia (Tosepu et al., 2020), Norway (Menebo, 2020) and also over the globe (Sobral et al., 2020; Wu et al., 2020) are controversial and The World Health Organization (WHO) has stated that more research should be focused on how to quantify how the weather affects the virus's spread.

7.2.4 Time Series Analysis to Predict COVID-19 Cases

It is clear from previous research that time series models such as exponential smoothing, ARIMA, and SARIMA performed well and provided adequate results for COVID-19 prediction. Many scholars have researched COVID-19 virus infection predictions. All previous research has established that the ARIMA model is the most effective for forecasting (Sahai et al., 2020; Jain et al., 2021; Murugesan et al., 2020; Sahai et al., 2020; Sulasikin et al., 2020; Mustafa and Fareed, 2020; Benvenuto et al., 2020). Sulasikin et al. (2020) used three approaches to predict the COVID-19 instances (Holt's method, Holt-Winters method, and ARIMA). Among the other models, the ARIMA model was deemed the best by the authors. Furthermore, Nguyen et al. (2021) demonstrated that the COVID-19 infection incidence could be effectively incorporated locally into a VECM with the COVID-19 hospital data to improve the existing forecast models and produce precise short-term forecasts and practical situation-based long-term trajectories.

7.2.5 Data and Variables

The data utilised for the study spans the months of March 10, 2020, to December 31, 2020. Kuwait Environment Public Authority provided statistics on air pollutants

(O_3 , SO_2 , NO_2 , CO , and PM_{10}) (K-EPA). Kuwait's Ministry of Health provided the daily COVID-19 cases (MOH). Information at the Kuwait's Ministry of Health website (<https://corona.e.gov.kw/en>) presented a summary of the daily COVID-19 cases in Kuwait that related to MOH data. All of the variables were converted to their natural logarithms before using the model in order to make additive and linear models make more sense (Nelder and Wedderburn, 1972).

7.2.6 Air Quality Index (AQI)

The Air Quality Index (AQI) is a numerical indicator of a region's air quality. The AQI scale has the range 0 to 500, with a higher AQI value indicating poor air quality and a lower AQI (< 100) signifying good air quality in a given area. AQI values were calculated using 24-hour average PM_{10} , 8-hour average CO and O_3 , and 1-hour average NO_2 and SO_2 levels in the current study. The maximum AQI observed for a city was used as the overall AQI. As mentioned in section 1.3.5 on page 11, the AQI calculation was explained and performed using equation 1.1 on page 13 and table 1.2 on page 13.

7.3 Results and Discussion

7.3.1 The Descriptive Statistics

Table 7.1 shows the descriptive statistics for the air pollutant variables. The mean values corresponding to O_3 , CO , PM_{10} , SO_2 and NO_2 were 24.82 ± 7.20 , 9.11 ± 3.61 , 79.51 ± 24.45 , 11.24 ± 5.21 and 26.72 ± 13.00 respectively. It is also evident that except O_3 , all the pollutants were positively skewed, i.e. the mean values of these pollutants were high as compared to the median values. Moreover the Shapiro-Wilk test shows that the distributions of the variables were significantly different from a normal distribution. Therefore log-transformation will be applied on the variables to convert the distribution of the variable to to make them more normally distributive, before performing any further analysis. Minimum, maximum and percentile values of the pollutants are also shown in Table 7.1.

Table 7.1: Descriptive statistics for air pollutants.

	O_3	CO	PM_{10}	SO_2	NO_2
Mean	24.818	9.112	79.511	11.239	26.719
Std. Deviation	7.202	3.611	24.452	5.214	12.996
Skewness	0.026	1.969	2.542	1.383	0.514
Kurtosis	0.514	6.324	11.931	2.325	0.521
Shapiro-Wilk (SW)	0.991	0.844	0.813	0.895	0.959
P-value of (SW)	0.066	< .001	< .001	< .001	< .001
Minimum	9.250	3.791	35.357	3.802	5.154
Maximum	42.226	28.445	234.057	34.443	64.515
25th percentile	19.249	6.618	67.104	7.523	15.502
50th percentile	25.398	8.184	76.334	10.001	24.529
75th percentile	29.755	10.936	87.004	13.906	35.561

Descriptive statistics for daily climatology variables (RH, Temp, WD and WS), COVID-19 cases and COVID-19 deaths are shown in Table 7.2. The mean values for RH, Temp, WD and WS are 35.51(S.D. = 20.06), 30.26(S.D. =7.96), 206.138(S.D. =54.48) and 2.18(S.D. =0.66) respectively. Moreover, on average, 506 cases of COVID-19 and 3 deaths due to COVID-19 were reported in the study period. Results of the Shapiro-Wilk test show that the distributions of the climatology parameters, COVID-19 cases and COVID-19 deaths were different from the normal distribution. Therefore log-transformation will be applied on the variables to convert the distribution of the variables to be more normal. The values of other statistics, i.e. median (50th percentile), skewness, kurtosis, minimum, maximum and percentiles, for each variable are also shown in Table 7.2.

7.3.2 Correlation Analysis

Table 7.3 presents results of the correlation analysis for air pollutants, climatology parameters and number of COVID-19 cases. A strong significant positive correlation was observed between temperature and COVID-19 cases ($r_p = 0.61$), indicating that as the value of temperature increases, number of COVID-19 cases also increases, whereas a negative significant correlation was observed between RH and COVID-19 cases

Table 7.2: Descriptive statistics for weather climatologies and number of COVID-19 cases (infected and death cases) in Kuwait.

	RH	Temp	WD	WS	COVID-19 Cases	Death Cases
Mean	35.506	30.255	206.138	2.183	506.436	3.145
Std. Deviation	20.062	7.956	54.481	0.662	276.988	2.524
Skewness	0.793	0.371	0.123	0.786	0.290	0.918
Kurtosis	0.367	1.069	1.202	0.106	0.902	0.513
Shapiro-Wilk (SW)	0.897	0.935	0.948	0.948	0.956	0.915
P-value of (SW)	< .001	< .001	< .001	< .001	< .001	< .001
Minimum	11.619	12.199	89.663	0.938	1.000	0.000
Maximum	91.534	43.150	297.845	4.271	1073.000	11.000
25th percentile	18.564	24.714	160.862	1.670	278.000	1.000
50th percentile	28.010	31.141	199.509	2.020	554.500	3.000
75th percentile	51.462	37.501	257.648	2.606	711.750	4.000

RH: Relative humidity, Temp.: Temperature in Celsius, WD: Wind direction, WS: Wind speed, COVID-19 Cases: daily reported cases from Kuwait ministry of health

($r_p = -0.49$), indicating that as the value of relative humidity (RH) increases, number of COVID-19 cases decreases. Moreover, a small effect of O_3 ($r_p = 0.25$), CO ($r_p = 0.24$), PM_{10} ($r_p = 0.18$) and NO_2 ($r_p = 0.22$) was also observed on COVID-19 cases.

Additionally to the correlation analysis, regression analysis has been performed to check how the pollutant and climatology variable are related to the COVID-19 cases. Table 7.4 demonstrates the results of the linear regression analysis. The model so formed by regression analysis was statistically significant ($F(9,286) = 29.16$, $p < 0.01$, $R^2 = 0.48$). The results reveal that the air pollutants O_3 ($\beta = 7.70$, S.E. = 2.41, $t = 3.20$, $p < 0.05$) and CO ($\beta = 20.38$, S.E. = 4.88, $t = 4.17$, $p < 0.001$) significantly and positively affect COVID-19 cases in Kuwait. It indicates that an augmentation in O_3 and CO by 1 unit increases the expected number of COVID-19 cases by 7.70 and 20.38 units respectively. Additionally, the climatology parameters, temperature ($\beta = 17.38$, S.E. = 2.73, $t = 6.36$, $p < 0.001$) and wind speed ($\beta = -58.36$, S.E. = 24.54, $t = -2.38$, $p < 0.05$) also have shown significant effects on COVID-19 cases. It implies that a one unit increase in temperature will increase the expected COVID-19 cases by 17.38 units, whereas, a one unit increase in wind speed will decrease the COVID-19 cases by 58.36 units on average.

Table 7.3: Pearson’s Correlation test between the number of COVID-19 cases and air pollutants in Kuwait during March 10, 2020, to December 31, 2020.

Variable	O_3	CO	PM_{10}	SO_2	NO_2	RH	Temp	WD	WS
1. O_3	–								
2. CO	-0.419***	–							
3. PM_{10}	0.124*	0.080	–						
4. SO_2	0.033	0.008	-0.044	–					
5. NO_2	-0.452***	0.614***	-0.002	0.132*	–				
6. RH	-0.491***	0.228***	-0.228***	-0.391***	-0.027	–			
7. Temp	0.461***	0.009	0.292***	0.124*	0.010	-0.753***	–		
8. WD	0.026	-0.233***	0.085	0.272***	-0.035	-0.411***	0.134*	–	
9. WS	0.350***	-0.333***	0.317***	-0.088	-0.492***	-0.139*	0.206***	0.317***	–
10. COVID-19 Cases	0.248***	0.235***	0.182**	0.114	0.222***	-0.487***	0.608***	0.157**	-0.025

* p < .05, ** p < .01, *** p < .001

Table 7.4: Regression Coefficients to estimate the influence from air pollutants toward the changes in COVID-19 daily cases.

Variable	Unstandardised	Standard Error	Standardised	t	p	95% CI	
						Lower	Upper
(Intercept)	-494.427	182.605		-2.708	0.007	-853.847	-135.006
O_3	7.699	2.408	0.200	3.197	0.002	2.959	12.438
CO	20.382	4.884	0.266	4.173	< .001	10.768	29.996
PM_{10}	0.122	0.545	0.011	0.224	0.823	-0.951	1.195
SO_2	-2.269	2.662	-0.043	-0.852	0.395	-7.510	2.971
NO_2	1.852	1.404	0.087	1.319	0.188	-0.911	4.616
RH	-0.344	1.337	-0.025	-0.257	0.797	-2.976	2.288
Temp	17.377	2.732	0.499	6.360	< .001	11.999	22.755
WD	0.991	0.285	0.195	3.483	< .001	0.431	1.551
WS	-58.359	24.539	-0.140	-2.378	0.018	-106.659	-10.059

Note. Results: $F(9,286) = 29.16$, $p < 0.001$, $R^2 = 0.48$

7.3.3 Results of Unit Root and Granger Causality test

The Granger Causality test has been conducted to check if the series of independent variables is useful for making prediction or not. Results of the ADF Root Test, KPSS, Phillips-Perron (PP), Lag selection criterion for the VAR model, and the Granger Causality test will be presented in the following subsection:

Results of the Unit-Root Tests

After acquiring the significant association results for some of the air pollutants and climatology parameters with COVID-19 cases, VECM analysis has been carried out to evaluate short and long term relationships of these variables with COVID-19. It is necessary to perform stationarity tests on the time series data before performing VECM analysis, as the non-stationary series may produce spurious regression results for VECM (Asari et al., 2011; Latief et al., 2021). Therefore, tests have been performed for

integrating properties of the series to check if the series are stationary at the original level or not. To do this, the conventional Augmented Dickey-Fuller (ADF), and Phillips-Perron (PP) and KPSS tests have been performed. Table 7.5 and Table 7.6 show the results of the ADF test. Results of the PP test and KPSS test are demonstrated in Table 7.7 and Table 7.8 respectively.

All variables are transformed in their log forms to mitigate inconsistency in the data and ease interpretation of the results via elasticities:

$$\ln COVID19_t = \beta_0 + \beta_1 \ln[O_3]_t + \beta_2 \ln[SO_2]_t + \varepsilon_t. \quad (7.1)$$

Table 7.5: ADF Root Tests with constant.

	level			First Difference		
	coefficient	t-ratio	p-value	coefficient	t-ratio	p-value
Log(O_3)	-0.0708452	-1.854	0.3548	-0.373708	-5.953	7.67e-09***
Log(CO)	-0.151600	-3.869	0.0023*	-0.120178	-1.911	0.0570
Log(PM_{10})	-0.401449	-5.633	8.68e-07***	-0.129774	-1.770	0.0778
Log(SO_2)	-0.235017	-4.358	0.0003***	-0.183925	-2.763	0.0061**
Log(NO_2)	-0.0768789	-2.393	0.1436	-0.253951	-4.118	5.01e-05***
Log(COVID-19 Cases)	-0.0460937	-2.932	0.0417*	-0.393289	-6.970	2.18e-11***

*Stationarity at 5% significance levels **Stationarity at 1% significance levels *** Stationarity at 0.1% significance levels

Table 7.6: ADF Root Tests with constant and trend.

	level			First Difference		
	coefficient	t-ratio	p-value	coefficient	t-ratio	p-value
Log(O_3)	-0.183827	-3.594	0.0303*	-0.305549	-4.686	4.31e-06***
Log(CO)	-0.247380	-5.165	8.51e-05***	-0.0613829	-0.9560	0.3399
Log(PM_{10})	-0.427096	-5.832	2.93e-06***	-0.112507	-1.518	0.1300
Log(SO_2)	-0.235853	-4.362	0.0025**	-0.183525	-2.752	0.0063**
Log(NO_2)	-0.184001	-4.066	0.0070**	-0.187636	-2.938	0.0036**
Log(COVID19 Cases)	-0.0382004	-2.171	0.5055	-0.403573	-7.036	1.47e-11***

*Stationarity at 5% significance levels **Stationarity at 1% significance levels *** Stationarity at 0.1% significance levels

The ADF root test with constant shows that the time series for pollutants CO , PM_{10} , SO_2 and the series of COVID-19 cases are stationary at the 5%-level, whereas the series of O_3 and NO_2 were stationary at the first difference (see table 7.5). The results of the ADF test with constant and trend demonstrate that the series of all the pollutants are

Table 7.7: Unit root tests using KPSS with constant and trend.

	level			First Difference		
	coefficient	t-ratio	p-value	coefficient	t-ratio	p-value
Log(O_3)	-0.00227119	-13.02	6.43e-31***	-8.82232e-05	-0.6322	0.5278
Log(CO)	0.00210608	10.56	2.53e-22**	6.37E-05	0.4359	0.6632
Log(PM_{10})	-0.000566757	-3.161	0.0017**	9.78E-06	0.05286	0.9579
Log(SO_2)	0.000286053	0.9786	0.3286	4.94E-05	0.2043	0.8383
Log(NO_2)	0.00402837	14.63	8.82e-37***	7.88E-05	0.3921	0.6953
Log(COVID19 Cases)	0.00685064	8.903	5.74e-17***	-0.000319900	-1.302	0.1940

*Stationarity at 5% significance levels **Stationarity at 1% significance levels *** Stationarity at 0.1% significance levels. The KPSS test using the trend option rejects the hypothesis of trend stationarity at the 5% significance level.

Table 7.8: Phillips-Perron (PP) Unit Root Test constant and trend.

	constant			trend		
	coefficient	t-ratio	p-value	coefficient	t-ratio	p-value
Log(O_3)	0.80646	22.523	2e-16**	6.76E-01	15.73	2e-16**
Δ Log(O_3)	-0.301562	-5.398	1.4e-07**	-0.3029678	-5.417	1.27e-07**
Log(CO)	0.32928	5.971	6.81e-09**	3.10E-01	5.567	5.87e-08**
Δ Log(CO)	-0.35475	-6.482	3.85e-10**	-3.55E-01	-6.471	4.13e-10**
Log(PM_{10})	0.49101	9.655	2e-16**	4.74E-01	9.203	2e-16**
Δ Log(PM_{10})	-0.26979	-4.784	2.73e-06**	-2.70E-01	-4.776	2.84e-06***
Log(SO_2)	0.66515	15.129	2e-16**	6.64E-01	15.066	2e-16**
Δ Log(SO_2)	-0.24294	-4.279	2.55e-05**	-2.43E-01	-4.274	2.61e-05**
Log(NO_2)	0.8495	27.27	2e-16**	7.28E-01	18.49	2e-16**
Δ Log(NO_2)	-0.184247	-3.198	0.00153**	-1.85E-01	-3.202	0.00152**
Log(COVID19 Cases)	0.9394	58.309	2e-16**	0.9354317	51.365	2e-16**
Δ Log(COVID19Cases)	-0.29767	-5.327	1.99e-07**	-0.306341	-5.489	8.81e-08**

*Stationarity at 5% significance levels **Stationarity at 1% significance levels *** Stationarity at 0.1% significance levels, The null hypothesis of non-stationarity is rejected in all the cases, this shows that the lagged series are stationary at 5% level of significance.

stationary at the original level, whereas the series of COVID-19 was stationary at the first difference (see table 7.6). The results of the PP unit root test show that the series of all the pollutant variables and COVID-19 cases are stationary at the original level as well as on the first difference (Table 7.8). The results of the KPSS test show that the series of all the pollutants (except SO_2) and COVID-19 cases are non-stationary at the original level, though all the series are found to be stationary at the first difference. Additionally, the plots of the sample autocorrelation function (SACF) and sample partial autocorrelation function (SPACF) for the residuals are shown in Figure 7.1. From the figures, there are spikes at lag 22 in both SACF and SPACF indicating significant correlation between residuals.

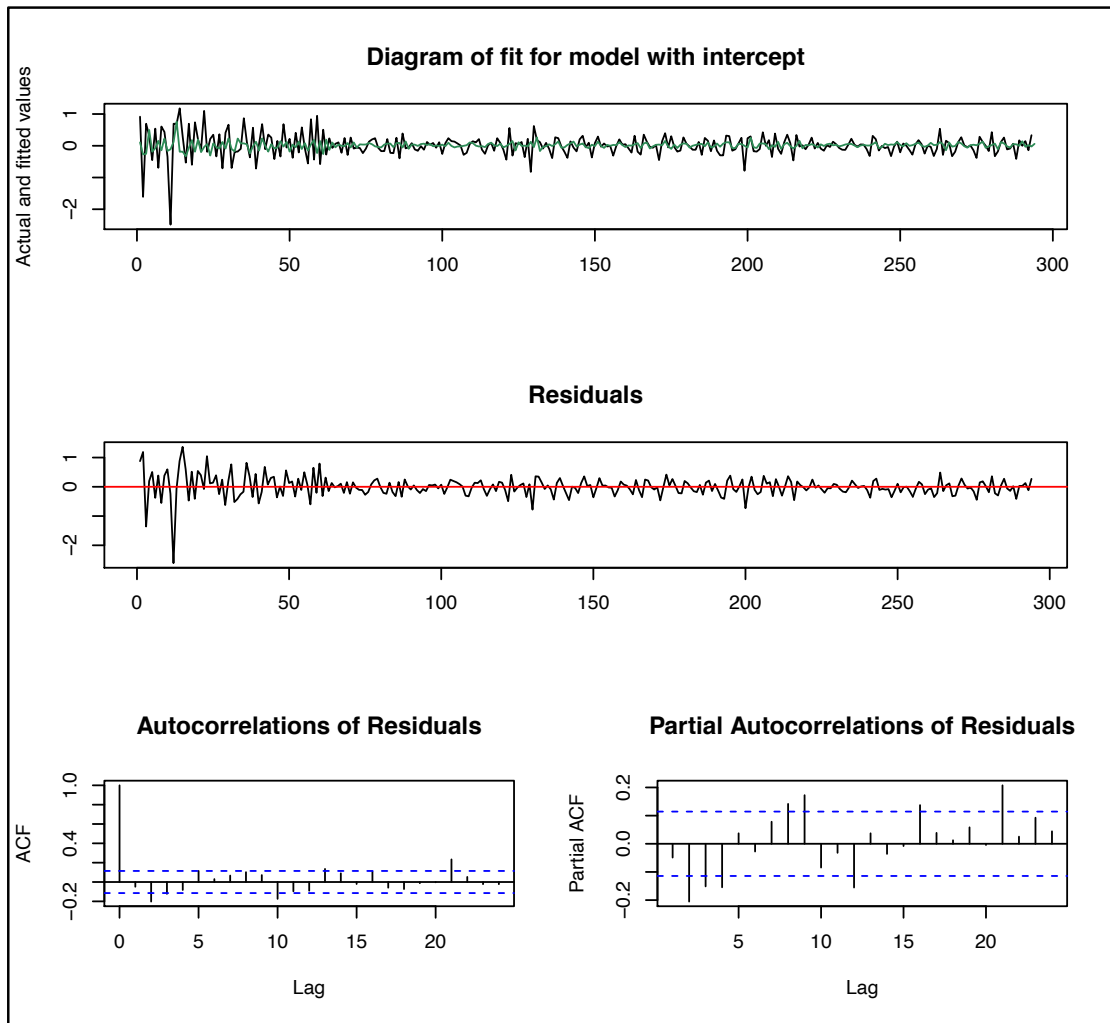


Figure 7.1: Log COVID-19 Kuwait daily cases, first difference, SACF and SPACF of the residuals of the multivariate model.

Estimation of VAR Model

After checking the stationarity of the series, the next step is to determine the number of optimal lags. To choose the number of lags that need to be included in the VAR model, the VARselect function has been taken into consideration. This function calculates four different information criteria across a number of different lags (up to a maximum specified within the function) and chooses the lag that has the lowest information criteria for each

of the four statistics. The asterisk symbol indicates the best values under the respective information criteria, AIC = Akaike criterion, BIC = Schwarz Bayesian criterion and HQC = Hannan-Quinn criterion. Table 7.9 and Table 7.10 illustrate the results of these lag order statistics. The Akaike information criteria statistics suggests that the optimal lag order for the model is 2.

Table 7.9: Lag selection criterion VAR Test using constant model with the endogenous series Log(COVID-19 Kuwait), Log(O_3), Log(SO_2), Log(NO_2), Log(CO).

lags	loglik	p(LR)	AIC	BIC	HQC
1	85.07777		-0.305516	0.236896*	-0.088003*
2	135.59486	0.00000	-0.408474*	0.598862	-0.004521
3	166.70389	0.00429	-0.373786	1.098474	0.216606
4	194.71554	0.01782	-0.317131	1.620053	0.459701
5	227.19868	0.00217	-0.292189	2.10992	0.671083
6	250.20871	0.12236	-0.200062	2.666971	0.94965
7	277.22468	0.02721	-0.136345	3.195612	1.199806
8	292.43136	0.73111	0.011125	3.808007	1.533716
9	316.40908	0.08782	0.09639	4.358196	1.80542
10	339.38759	0.12365	0.18874	4.915471	2.084211

*Stationarity at 5% significance levels ** Stationarity at 1% significance levels.

Table 7.10: Lag selection criterion VAR Test using trend model with the endogenous series Log(COVID-19 Kuwait), Log(O_3), Log(SO_2), Log(NO_2), Log(CO).

lags	loglik	p(LR)	AIC	BIC	HQC
1	119.82876		-0.520498	0.109137*	-0.267836*
2	172.34973	0.00000	-0.640215*	0.461646	-0.198057
3	196.71379	0.07647	-0.555897	1.01819	0.075758
4	217.67414	0.22951	-0.446914	1.599399	0.374237
5	241.60391	0.08932	-0.359449	2.159091	0.651199
6	270.35655	0.01286	-0.306932	2.683834	0.893213
7	299.06562	0.01311	-0.254099	3.208893	1.135542
8	317.95827	0.38768	-0.130132	3.805086	1.449005
9	346.78133	0.01246	-0.078126	4.329319	1.690508
10	370.4986	0.09622	0.01088	4.89055	1.96901

*Stationarity at 5% significance levels ** Stationarity at 1% significance levels.

Johansen Cointegration Tests

The Johansen cointegration test has been performed to check the long term relationship among the variables. In order to explain the long term relationships among the variables is necessary to conduct VECM analysis. The Johansen cointegration test fails to reject the null hypothesis saying there is no level of cointegration ($r = 0$, trace test=130.71, $p = 0.00$). This reveals that there exists at least one level of cointegration equation, which indicates that the variables have a long-term relationship. Further, the results of the cointegration test show that there exist at most two levels of cointegration ($r \leq 1$, trace test=83.29, $p=0.00$) between the times series of Log(COVID-19 Kuwait), Log(O_3), Log(SO_2), Log(NO_2), Log(CO) (Table 7.11). Overall, it can be concluded that there is a long-term stationary equilibrium between the daily admitted COVID-19 cases and air pollution levels.

Moreover, Engle and Granger (1987) suggested a two step process to test the cointegration (an OLS regression and a unit root test). According to Engle and Granger (1987), if a set of variables are cointegrated, then there exists a valid error correction representation of the data, and vice-versa. Therefore, an analysis of OLS regression and the error correction model has been performed to test the cointegrating relationship ($r = 2$) in a system of $k = 2$, $I(1)$ variables. The results of cointegration regression analysis (Table 7.12) and the error correction model (Table 7.13), confirm that there is a long term relationship between the series Log(COVID-19 Kuwait), Log(O_3), Log(SO_2), and Log(CO). Figure 7.2 shows the simultaneous variation of Log(COVID19) with Log(CO), Log(NO_2), Log(O_3), Log(SO_2), Log(PM_{10}) and Log(Humidity).

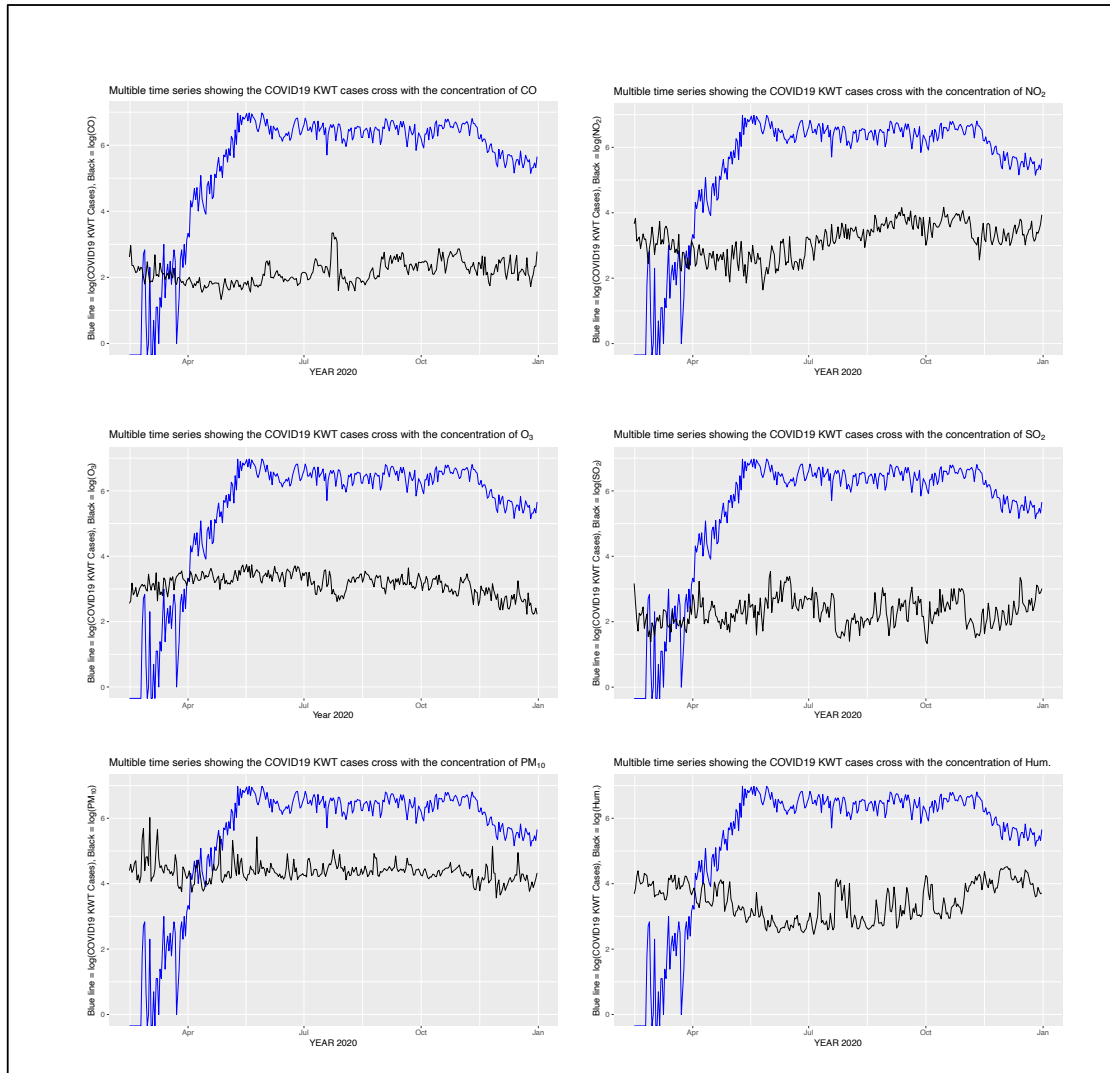


Figure 7.2: Daily time series for Log COVID-19 (Kuwait) compared with $\text{Log}(O_3)$, $\text{Log}(SO_2)$, $\text{Log}(NO_2)$, $\text{Log}(CO)$ and $\text{Log}(PM_{10})$.

7.3.4 Determination of Optimal VAR

Table 7.14 and Table 7.15 show the results of VECM analysis with restricted constant and restricted trend respectively. The analysis demonstrated that the past COVID-19 cases, and the pollutants O_3 and SO_2 significantly affect the future COVID-19 cases in Kuwait. The value of EC1 is negative and significant, which indicates the existence of a

Table 7.11: Johansen Test for selecting the best cointegration rank (r) that reflects linear combinations of underlying series to form a stationary series for Log(COVID-19 Kuwait), Log(O_3), Log(SO_2), Log(NO_2), Log(CO) and Log(PM_{10}). The asterisk reflects the best cointegration rank which is at $r = 1$.

Rank	Eigenvalue	Trace test	p-value	Lmax test	p-value
$r = 0$	0.14989	130.71	0.0000	47.419	0.0041
$r \leq 1^*$	0.1231	83.29	0.0023	38.359	0.0105
$r \leq 2$	0.06056	44.931	0.0907	18.242	0.4872
$r \leq 3$	0.047901	26.689	0.1122	14.333	0.3519
$r \leq 4$	0.031223	12.356	0.1417	9.2624	0.2710
$r \leq 5$	0.010538	3.0933	0.0786	3.0933	0.0786

Table 7.12: Cointegration regression for the series Log(COVID-19 Kuwait), Log(O_3) and Log(SO_2) with constant and trend.

	coefficient	std. error	t-ratio	p-value
const	-3.21775	0.824189	-3.904	0.0001**
$\Delta \text{Log}(O_3)$	2.09399	0.224713	9.318	2.99E-18***
$\Delta \text{Log}(SO_2)$	0.307531	0.134071	2.294	0.0225*
time	0.0115185	0.000844428	13.64	4.09E-33***

Dependent variable is Log(COVID-19 Kuwait), * Stationarity at 5% significance levels ** Stationarity at 1% significance levels *** Stationarity at 0.1% significance levels. R-squared = 0.409, Adjusted R-squared = 0.403, Akaike criterion = 833.977.

long-run causality of the future COVID-19 cases with the past COVID-19 cases.

Sometimes it is difficult to directly observe the relations between the variables in a VAR model from the parameter matrices. In that case, Impulse Response Functions (IRF) have been implemented as a tool for interpreting the VAR model (Lütkepohl, 2010). The impulse response function shows the changes in a variable over a period of time, when a shock is given to the other variable. In IRF, a shock was given to the pollutant variables CO , SO_2 , O_3 and COVID-19 and its impact was observed on COVID-19 (Figure 7.3) through the fitted VAR model. Figure 7.3 illustrates that a shock to the SO_2 and O_3 impacts the future value of COVID-19 in a positive manner in the longer run period. It is difficult to conclude directly from Figure 7.3 whether the

Table 7.13: Estimates from the Error Correction Model for the series Log(COVID-19 Kuwait), Log(O_3) and Log(SO_2) with constant and trend.

Residual	coefficient	std. error	t-ratio	p-value
uhat_1	-0.137586	0.0315632	-4.359	0.0256*
d_uhat_1	-0.186812	0.056996	-3.278	0.0012**

AIC: 437.476 BIC: 444.843 HQC: 440.426,* Stationarity at 5% significance levels ** Stationarity at 1% significance levels *** Stationarity at 0.1% significance levels. Stationarity test statistic: tau_ct(3) = -4.35906, asymptotic p-value 0.02562, 1st-order autocorrelation coeff. for e: -0.029.

Table 7.14: VECM Equation d_1_KWT_Cases with restricted trend, lag order = 2, cointegration rank order = 3.

	Coefficient	Std. Error	t-ratio	p-value
const	-0.016889	0.318250	0.053	0.9577
$\Delta \log(COVID19)$	0.273270	0.055023	4.966	<0.0001 ***
$\Delta \log(O_3)$	0.242484	0.108913	2.226	0.0268 **
$\Delta \log(SO_2)$	0.124354	0.062629	1.986	0.048**
EC1	-0.073499	0.019090	-3.850	0.0001 ***
EC2	0.090969	0.099738	0.912	0.3625
EC3	0.061642	0.051422	1.199	0.2316
Mean dependent var	0.015501	S.D. dependent var		0.360049
Sum squared residuals	32.03256	SS.E. of regression		0.334667
R-squared	0.156664	Adjusted R-squared		0.136023
rho	-0.030311	Durbin-Watson		2.046586

changes in impulse response functions are significant or not, as the error margin (shown with dotted lines) is very high. Moreover, a current shock to the CO and COVID-19 indicate no long run impact on the future values of COVID-19.

Figure 7.4 illustrates the result of Forecast Error Variance Decomposition (FEVD). It can be observed that COVID-19 itself influences the future error forecast variance of COVID-19. Moreover, a small effect of O_3 can be observed on the future error forecast variance of COVID-19. The effects of CO and SO_2 have not shown any influence on COVID-19 cases. Therefore, it can be concluded that the future value of COVID-19 will only be affected by O_3 and past cases of COVID-19 itself.

Table 7.15: VECM Equation $d_1_KWT_Cases$ with restricted constant, lag order = 2, cointegration rank order = 3. Note that: * $p < .05$, ** $p < .01$, *** $p < .001$.

	Coefficient	Std. Error	t-ratio	p-value
$\Delta \log(COVID19)$	0.278729	0.054523	5.112	< 0.0001 ***
$\Delta \log(O_3)$	0.222039	0.103087	2.154	0.0321 *
$\Delta \log(SO_2)$	0.126831	0.062415	2.032	0.0431 *
EC1	-0.065624	0.015564	4.216	< 0.0001 ***
EC2	0.046889	0.065545	0.715	0.475
EC3	0.062704	0.051343	1.221	0.223
Mean dependent var	0.015501	S.D. dependent var		0.360049
Sum squared residuals	32.07109	SS.E. of regression		0.334284
R-squared	0.157217	Adjusted R-squared		0.139598
rho	0.031156	Durbin-Watson		2.038010

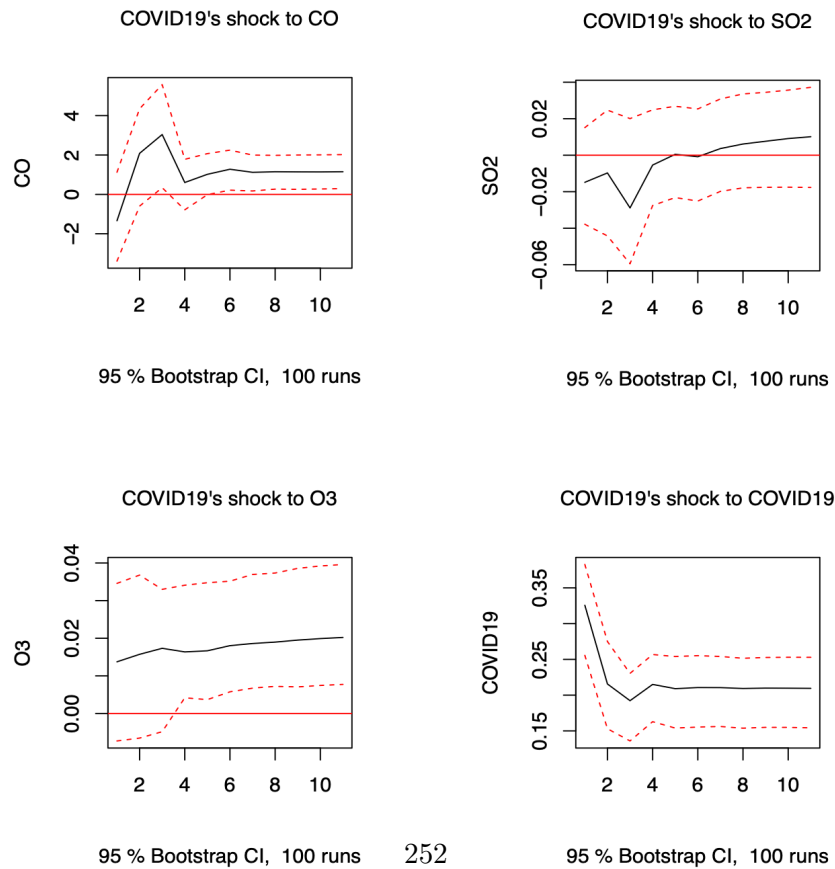


Figure 7.3: Impulse Responses

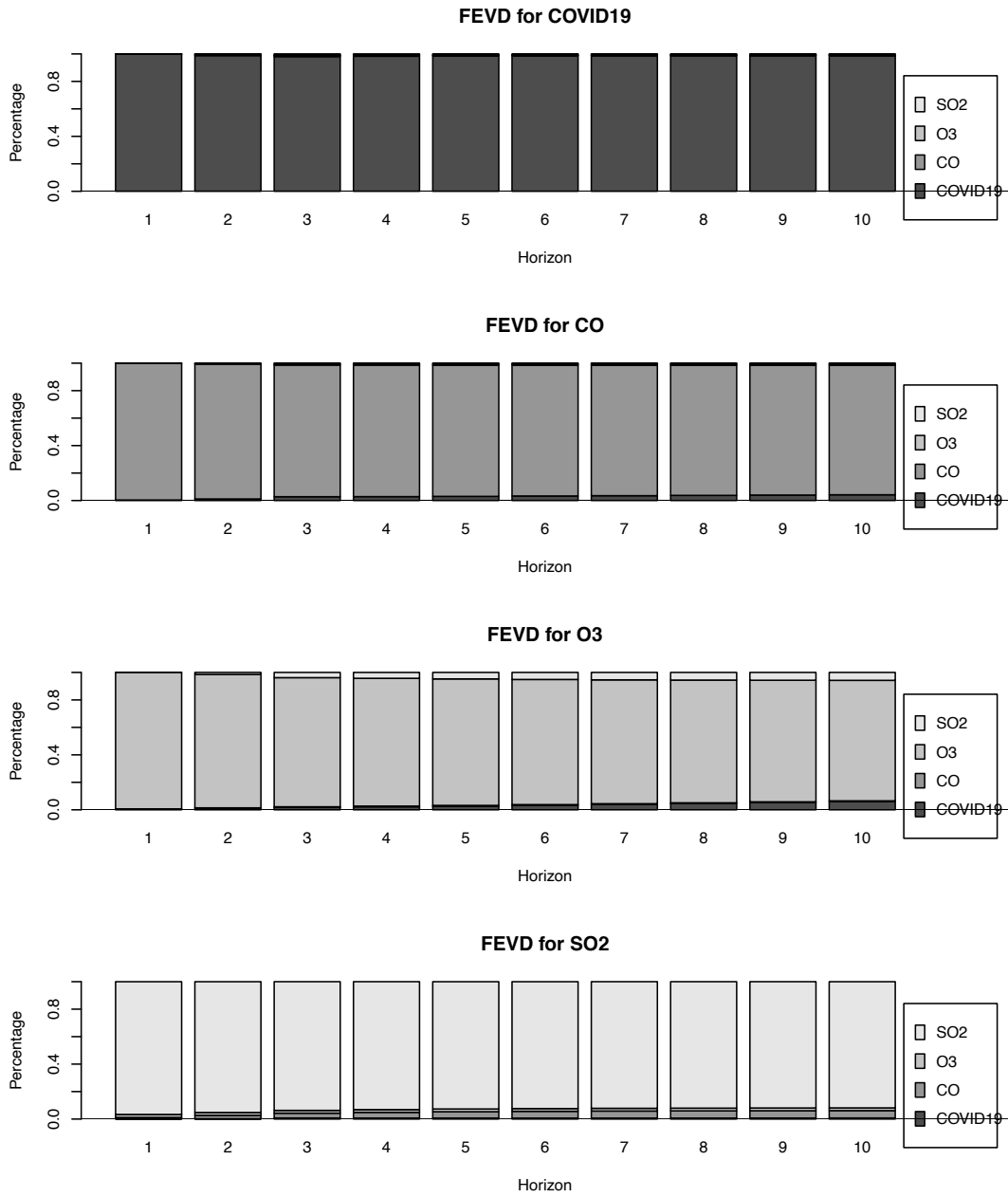


Figure 7.4: Forecast Error Variance Decomposition (FEVD).

Figure 7.5 shows the future trend of COVID-19 by using VECM. From Figure 7.5, it can be observed that the forecasted value shows a linear trend and the predicted value

lies within the 95% confidence interval.

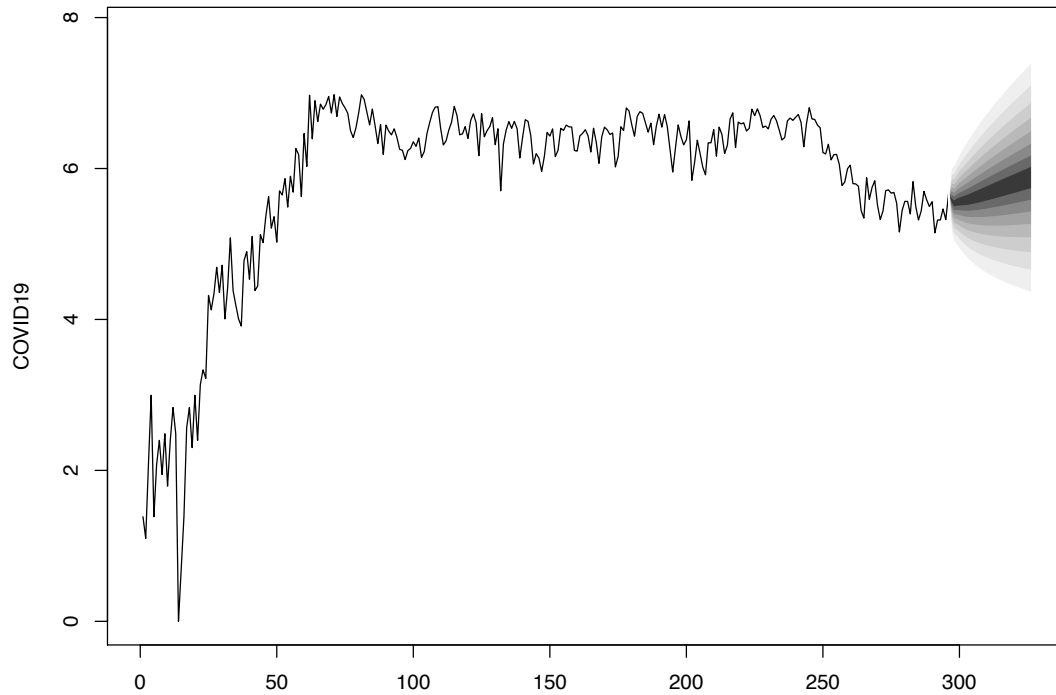


Figure 7.5: Forecasting COVID-19 cases using VECM.

7.4 Conclusion and Recommendations

The primary goal of the current study is to look into the association between changes in daily admitted COVID-19 cases and air pollution levels during the Corona pandemic from March to December 2020. Based on a descriptive analysis of the variables, the association between air pollutants (O_3 , SO_2 , NO_2 , CO , and PM_{10}) and daily admitted COVID-19 patients has been established, and this is consistent with the literature reviewed. This research used the vector error corrected model (VECM) with the cointegration technique to look at the long and short run association between the effect of air pollution (O_3 , SO_2 , NO_2 , CO , and PM_{10}) and the daily admitted COVID-19 cases. We discovered that for COVID-19 patients, a greater AQI was linked to a higher number of

hospitalisations.

When the health variables were examined, it was shown that the majority of the people infected with COVID-19 were already exposed to air pollution because Kuwait's regions have significant pollution rates. The biggest cause of pollution has been air pollutants emitted by cars and businesses (Hamoda et al., 2022). COVID-19 impacts the human respiratory system, and people who are already susceptible to respiratory disease have a propensity to be affected by the pandemic (Domingo and Rovira, 2020).

COVID-19's lockdown paralysed human activities, mostly involving vehicle usage and public transportation, as well as industrial processes (Pata, 2020; Gautam, 2020; Shehzad et al., 2020). The importance of air pollution and COVID-19 has been demonstrated in numerous researches. The spread of COVID-19 has been found predominant through airborne bio-aerosol droplets together with various aspects of urban air pollution (Fareed et al., 2020). Past exposure to air pollution has led to an increase in the cases of COVID-19. The ability to transfer these viruses is demonstrated by air pollution. We approximated the error correction model based on the VECM procedure to obtain short-term coefficients after investigating the long-term findings. The results show that while O_3 and SO_2 have an increasing short-term effect, they have a long-term positive effect on the daily admitted COVID-19 cases. The error correction term (ECT) is statistically significant and has a negative value, indicating that a deviation from the long-term equilibrium will be repaired. The findings show that the short-term coefficients of O_3 and SO_2 are lower than the long-term coefficients.

Our research has several limitations. We have to revert to the air quality index as a measure of air pollution level due to inadequate reporting on certain pollutants. This, however, may obscure the impact of certain contaminants on the number of hospitalisations. Furthermore, because our estimates focused on a single link between factors, any ascribed cost estimation should be cautiously approached. Other aspects, such as humidity, wind speed, and seasonality level, may need to be adjusted in the model (winter, autumn, spring and summer). However, because their data was not available or valid in this study, we did not adjust for them.

Other time series methods, such as the vector autoregression (VAR) model, which is one of the most effective, flexible, and user-friendly models for multivariate time series analysis, could be recommended for future investigations. The basic model for studying a stationary time series in terms of two polynomials is the autoregressive-moving average (ARMA) process. Other multivariate time series analysis techniques include Vector Autoregression Moving-Average (VARMA), VARMAX (VARMAX with Exogenous Regressors), and Holt Winter's Exponential Smoothing (HWES). A spatial multivariate time series approach could be used to assess the distance between the location of a job or a living area and a pollution source.

Chapter 8

Discussion and Conclusion

8.1 Overview

Air pollution is presently the world's single, largest, environmental health risk (WHO, 2014a). Furthermore, air pollution and climate conditions influence the trends of viral respiratory illness epidemics by altering host immunity and virus survival period (Mirsaeidi et al., 2016). It has long been established that the exposure of atmosphere particles to ambient air pollutants (O_3 , SO_2 , NO_2 , CO , and PM_{10}) is closely connected to the incidence and mortality rate of respiratory disorders (WHO, 2014c; as cited in: (Ferreira et al., 2016; Li et al., 2021; Sacramento et al., 2020; Rodríguez-Camargo et al., 2020)). These elements have also been identified as the most important environmental predictors of sicknesses such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS) as well as premature death in humans.

Studies in a variety of contexts using an array of different methods and approaches have consistently shown that air pollution correlates with a wide range of health effects. This thesis contains a collection of papers covering several research topics with differing strategies for analysis and designs. Individual papers discuss the findings of the investigations in depth within the context of the contemporary literature. Difficulties in interpreting, conducting, analysing, and even presenting studies of air pollution's health effects provide a perspective for the discussions. An explanation of the well-known

approach for modelling multivariate time series analysis using a vector error corrected model to examine the relationship between air pollution and health outcomes over time provides additional perspective.

8.2 The Characteristics of Air Pollution in Kuwait

In the first part of this thesis we explored the characteristics of air pollution in Kuwait from 2012 to 2017 using hourly data from ten fixed stations distributed throughout Kuwait. Specifically eight air pollutants were of interest: O_3 , NO , NO_2 , NO_x , SO_2 , CO , $NMHC_s$, and PM_{10} . The results of various statistical tests regarding the measurements of these pollutants led us to formulate the following conclusions:

- The daily SO_2 , NO_2 and PM_{10} concentrations exceeded the corresponding thresholds or permissible limits defined by the K-EPA in the residential areas.
- The comparison results for the industrial areas indicated significant differences among the air pollutants. The daily SO_2 , NO_2 and NO_x concentrations exceeded the K-EPA standard values only in the SUK area, whereas PM_{10} concentrations exceeded the K-EPA threshold value at all industrial sites.
- The concentrations of all pollutants in residential areas resulted from high levels of industrial and vehicular emissions in nearby areas and depended on meteorological conditions (PM_{10} and NO_x).
- A strong interdependence occurred between NO_x (NO and NO_2), indicating the high oxidation reaction. Relatively high correlation occurred between climatology variables (e.g., temperature and humidity) and air pollutants (e.g., O_3 and CO). Increase in ozone levels could lead to more respiratory illnesses.
- Concentrations of different pollutants varied by seasons, where the NO , NO_2 and NO_x concentrations were very high in winter, while the O_3 concentrations were high during the first intermonsoon season, peaking in summer (July).

The State of Kuwait faces a growing risk of health-related problems due to the poor air quality originating from its various industrial and domestic activities (Al-Hurban et al., 2021). Dust from adjacent deserts collects both living (biogenic) and non-living (chemical) constituents as it passes through areas with industrial emissions. Regular monitoring and careful statistical examination of all measured air pollutants could help in maintaining a clean healthy environment and resolving pollution-related problems in a timely manner.

8.3 The Relation Between RA and Ambient Air Pollution

In the second part of this thesis we demonstrated and explained the relationship between ambient air pollution and patients' Rheumatoid Arthritis (RA) Disease Activity Scores using an index with 28 joints (DAS28) and the Clinical Disease Activity Index (CDAI). With this knowledge, we used Generalised Additive Models (GAMs) to estimate the risk of developing RA due to air pollution (Alsaber et al., 2020). Our findings have significant implications on the relationship between increased concentrations of SO_2 and NO_2 air pollutants and RA disease activity, with lesser implications on that of O_3 and CO .

Particularly regarding SO_2 and NO_2 , our results show that increased concentrations of these pollutants may increase the risk of developing RA disease activity, as measured by both the *DAS28* and *CDAI*. These results offer significant support to those of other studies regarding SO_2 and NO_2 (Hart et al., 2013b,a).

Regarding O_3 , and CO , our results using the *DAS28* do not support those of a study from Korea that found an increased risk for developing RA in adults exposed to increased O_3 and CO concentrations (Shin et al., 2019). However, our results offer partial support when using the *CDAI*, which do indicate that O_3 in addition to SO_2 and NO_2 as reported above, but not CO , is a statistically significant risk factor for predicting RA disease activity.

Regarding PM_{10} , our results agree with the consensus of the literature that has not established a link between PM_{10} and RA disease activity (Hart et al., 2013b,a; Shin

et al., 2019). This reflects our results using both the DAS28 and CDAI measurements.

Generally, the major sources of air pollutants in Kuwait are oil refineries, traffic, and power plants, mostly using fossil fuels. These are also thought to be the main sources of SO_2 and NO_2 in Kuwait City (Al-Awadhi, 2014). Additionally, the most prevalent pollutants generated from road transport are NO_2 , CO, SO_2 , volatile organic compounds (VOCs), and particulate matter (Hankey et al., 2012).

With a population of over four million people and a fleet of more than two million vehicles both growing rapidly, Kuwait is experiencing increases in traffic volume, trip frequency, and trip length (Al-Mutairi et al., 2009), causing the poor air quality to become an increasingly major concern, especially for people living in Kuwait City. Kuwait is a relatively small country for such a high quantity of fixed and especially mobile sources of different pollutants affecting air quality. Automobiles are particularly responsible for VOCs and NO_2 emissions, whereas power stations and water distillation plants using fossil fuel combustion to support Kuwait's oil exports are liable for the elevated atmospheric SO_2 concentrations (Al-Awadhi, 2014).

8.4 Dealing with Missing Data

Missing data, on the other hand, can cause a variety of issues in statistical modelling. To begin with, the lack of data diminishes statistical power, which refers to the likelihood that the test will reject the null hypothesis if it is wrong. Second, missing data can lead to parameter estimation bias. Third, it may reduce the sample's representativeness. Fourth, it may make the study's analysis more difficult. Each of these inaccuracies can jeopardise the trial's legitimacy and lead to erroneous results (Kang, 2013), therefore it is important to consider factors that may potentially facilitate missing data before beginning research (Graham, 2012). In doing so, researchers can measure these factors influencing data missingness and do extensive analysis.

8.4.1 Dealing with Missing Data - Air Pollution Data

In chapter 4 we presented several advanced statistical approaches relating to deep learning machines prepared for treating missing observations. Many contributions have been made to this discipline, such as in environmental science (Junninen et al., 2004; No-razian et al., 2008; Plaia and Bondi, 2006; Kabir et al., 2019), statistics (Di Zio et al., 2007; Huisman and Krause, 2018), and medicine (Sartori et al., 2005; Branden and Verboven, 2009). In environmental science imputation refers to a statistical process for giving inferential figures to all missing data by utilising prior information from other factors. Once a nascent process, as more people have become knowledgeable about imputation algorithms, inquisitiveness regarding the methodology has increased, leading to the invention of more and more sophisticated imputation methods. The existence of these elaborate and efficient imputation algorithms has led to their extensive usage around the world. However, the main challenge concerning imputed values is whether to consider them as actual measurements, or to be handled with caution. In the field of research, it is preferable to handle imputed data with great discretion. The use of imputed figures as actual data may misguide researchers into potentially falsifying the final results. Therefore, imputed values should be given low priority, and it is vital for a researcher to assess the robustness of the associated data estimation when working with imputed data. However, environmental information that relies on technological processing and simulation remains a challenge.

Another approach to treating missing data is data ascription. A substantial quality of ascription methods is that they are reliable and limited to one type of variable. This variable may be considered as persistent or unmitigated. If the data type is mixed, the method must deal with the different types of data separately. Ultimately, while reliable for single-factor variables, these techniques ignore the potential associations between different factor types. To avoid any bias and maximise model performance for accurate estimation, it is essential when working with multi-factor variables to treat the missing values and estimate them using information from other predictors before conducting any statistical modelling or performing a time series analysis.

A third approach to treating missing data is the missForest imputation method, which has a consistent and comparatively lower imputation error (of 0.82) (Alsaber et al., 2021b). The approach had a minimum root mean square error (RMSE) equal to 1.04. MissForest also exhibited the smallest prediction deviation for the imputed values of pollutants. Furthermore, missForest simulation provides the most readily available imputation method for missing values, as its freeware R package is freely available.

When compiling our report of the study, we utilised the missing at random (MAR) tool, which assumes values of missing variables are not necessarily directly related to their causes. This premise is essential for the development of a prototype of the observation for the imputation of missing data. There was a possibility that the missing data was not missing at random (MNAR). In such a case, the values of missing variables are directly related to their causes. Distinguishing between MNAR and MAR would involve a meticulous investigation of the data capturing process, and, if determined to be MNAR, it can still be challenging to determine the actual missing data mechanism. Other potential assumptions include Gaussian-distributed data, which also could have been erroneous for some variables.

8.4.2 Dealing with Missing Data - KRRD Data

The process of estimating the missing data was repeated on data from the Kuwait Registry for Rheumatic Diseases (KRRD). We came to the same conclusion: MissForest is a very accurate approach of missing data imputation in KRRD that surpasses other standard imputation strategies, in terms of both minimising imputation error and maintaining predictive performance in clinical models. This method can be used to maximise the usefulness of data in registries, such as those for patients with RA (Alsaber et al., 2021a). In general, our results show that MI using the MAR mechanism had the lowest RMSE among the other missingness mechanism (MCAR or MNAR). Compared with Complete Case Analysis (CCA), its effectiveness is due to MI's use of information in incomplete cases, while CCA is only valid in the case of MAR or MCAR data (Little, 1992). In well-designed studies, such as clinical trials, MAR mechanisms are more

common than MCAR, because in most cases, observable data explain most of the deficiencies (Molenberghs and Kenward, 2007). However, the MI technique sometimes presents challenges that prevent it from being the better method, even when the MCAR or MAR mechanism holds. For example, a small sample size may minimise the accuracy of MI (McKnight et al., 2007). Additionally, the utilisation of MI in longitudinal designs with layered data may necessitate the use of alternate approaches (Enders, 2010, 2011; Graham, 2012). Another challenge is that statistical packages vary in how easy it is to merge variables and to conduct test statistics because sometimes this requires programming knowledge. Even so, by providing numerous missing data methods, MI tends to have the most benefit in the clinical surrounding.

Additionally, it also indicates the significance of having a comprehensive understanding of the type and the effect of the missing data despite active handling of the data to manage and estimate the missing values. It is also important to consider factors that may potentially cause missing data before the beginning of the research (Graham, 2012). That way, researchers can measure these factors influencing data missingness and do extensive analysis.

8.5 Time Series Modelling to Predict COVID-19 Cases using the Information of Air Pollution

The main objective of this thesis has been to build a multivariate time series model to measure the short- or long-term relationship between the effect of air pollutants (SO_2 , NO_2 , CO , O_3 , and PM_{10}) with chronic disease in Kuwait. As mentioned in chapter 1, we first applied a multivariate time series Vector Error-Correction Model (VECM) to the number of admitted hospital patients with COVID-19 to measure the causal relationship of air pollution on the total number of hospital admissions. Secondly, we applied a VECM multivariate time series model to forecast the short- and long-term association between air pollution and patients with rheumatoid disease.

The VECM results showed that increasing the air quality index (AQI) had a signifi-

cant and positive effect on increasing the admitted number of COVID-19 patients over time, as with increased pollution came increased hospital admissions. The effect was not immediate, however. The lags of the dependent variable showed significance until the second lags, implying that there were long delays between an increased AQI particularly for O_3 and SO_2 and an increased number of COVID-19 patients. The model revealed that a shock intended to achieve long-term balance rectified approximately 7% of the short-term imbalance in just one day. In the long term, boosting the AQI for O_3 and SO_2 successfully increased the number of admitted COVID-19 patients.

The coefficient of the air pollution index was positive and significant for modelling the number of hospitalised COVID-19 patients. This suggests that raising the AQI of O_3 and SO_2 can increase the number of COVID-19 patients admitted to hospital. This finding supports another study that found air pollution to have a significant impact on COVID-19 infection and mortality rates (Frontera et al., 2021), and a third that used a time series method to find upsurges of approximately 6% in daily COVID-19 associated deaths daily to be significantly associated with large IQR increases in CO , NO_2 , and $PM_{2.5}$ (Dales et al., 2021).

8.6 Time Series Modelling to Predict RA Disease Activity Score (DAS28) using the Information of Air Pollution

Chapter 6 examined the linkage between SO_2 , NO_2 , O_3 and DAS28 scores for patients with RA in Kuwait using the Granger causality test and the Impulse Response Functions (IRFs) analysis. The Granger causality test analyses static causality using several time series approaches, including the VECM; while the IRFs analyse dynamic causality. We also utilised a comprehensive conceptual framework, including a panel VECM, cointegration test, and unit root test. The panel VECM provided data on the long-term causation and asymptotic convergence among the variables. Our empirical outcomes showed that NO_2 and O_3 are statistically significant in most of the study locations (ASA, FAH, MAN and JAH) when the DAS28 was the dependent variable. The results

demonstrate that the lagged Error Correction Term (ECT) coefficients in DAS28 and air pollution emissions are statistically significant.

The results of the Granger causality test reinforced the hypothesis that NO_2 and O_3 have a Granger causal relationship with DAS28 scores for RA in several residential locations in Kuwait (Table 6.11), though the study did not discover evidence of Granger-causality between SO_2 and DAS28 in VECM models. However, the study did discover that NO_2 and O_3 emissions in the environment have a Granger causal relationship with DAS28 scores in the long term. These results agree with expectations of factory emissions and are but an example of what is common to many underdeveloped and industrialising countries (Hwang and Yoo, 2014).

The Johansen cointegration test gives long-term associations to explain the relationship between air pollutants and DAS28, and it also provides the magnitude of correction needed for long-term deviations. If the system assembles to a long-term equilibrium from the short-term changes, the magnitude of the error correction term is also established for the time series SO_2 , NO_2 , and DAS28. We employ cointegration and vector error correction modelling to identify functional correlations, which we treat as endogenous variables in our analysis. For the unit root test, we first utilised the Dicky-Fuller Cointegration Regression test to rule out the existence of a unit root in the residuals. The test statistic was within the acceptable region for the null of no unit root. Next, we determined the number of cointegrating vectors using the Johansen trace test. For the rank of the coefficient matrix, not only the null of zero was rejected, but also rank one and two, instead, revealing that the rank should be three.

Although the VECM had confirmed the short- and long-term relationship of the DAS28 scores and the pollutants of SO_2 , NO_2 , O_3 , this data has the potential of inadequately or only very feebly supporting the second integration relationship. To supplement the data from the VECM, we also utilised a vector autoregressive (VAR) model to analyse cointegration in the short- and long-term associations between SO_2 , NO_2 , O_3 emissions and the DAS28. Consequently, we anticipate that our integrated analytical approach increases the relevance of these studies to better understand the interaction

between air pollution and rheumatoid arthritis disease progression.

The rise in the AQI for SO_2 , NO_2 and O_3 has a significant and positive impact on the DAS28 in RA patients, and the lags of the dependent variable are significant until the seventh lags, indicating that the increase in AQI affects patients, with extended delays. Other elements like temperature, relative humidity, and windspeed do not appear to have any impact on this relation. The error correction model revealed that in the event of a shock, 33% to 37% of the short-term imbalance is rectified in order to achieve long-term balance in just one day.

By comparison, in our previous study (Alsaber et al., 2020), we used numerous regression models to explain the link between air pollutants and RA disease activity, verifying that the link remained even after the addition of other highly significant RA covariates for CDAI and DAS28 scores. Those findings supported other studies such as Hart et al. (2013b) who explored the bearing of protracted exposure to air pollution on the possibility of suffering RA, in the Swedish Epidemiological Investigation. While there was no evidence of a link between PM_{10} exposure and an increased risk of RA in that study, and while the overall risks of gaseous pollutants on RA were somewhat elevated, they were not statistically significant after controlling for the variables of education and smoking.

In contrast, the results of this thesis established significant implications of a greater risk of RA incidence following elevated levels of NO_2 and SO_2 . However, the negative, one-lag error correction terms demonstrated a cointegrated association of the model's variables, meaning the variables had a different impact on DAS28 scores in the short-term as compared to the long-term. Although O_3 and NO_2 have a negative impact on DAS28 in the long-term, they have a favourable impact in the short-term, particularly so in the ASA and MAN locations. Conclusively, the VECMs confirm that for both short- and long-term equilibrium, O_3 and NO_2 both have causal effects on the DAS28 for RA patients who are living in residential areas surrounded by the sources of these pollutants.

8.7 Conclusion of the Study

8.7.1 Limitation of the Study

This study has several limitations. Firstly, there was a very limited number of published articles discussing the relation between air pollution and rheumatology disease using the CDAI and DAS28, so we have little with which to compare our findings. This, however, gives our study additional value, as future research can contribute to the literature by putting our findings to further test. Secondly, we acknowledge that two significant drawbacks of our study are that we did not measure $PM_{2.5}$ which has well-documented links to health hazards and that more than 60% of our total daily recordings had missing values. However, the latter allowed us to test various VECMs to provide additional data regarding their reliability and limitations to the available literature. Concerning the former, additional studies should take particle matter into account to help ascertain a more well-rounded perspective of our findings. Thirdly, in order to establish a link between specific RA patient's records and the data on air pollutants where they live, we encountered difficulties in linking some of them due to patients who have multiple addresses. Because of this, we were forced to exclude them from the study, limiting our sample size. However, we believe that we maintained a robust sample of RA patients within our study, and that it is approximately indicative of the population of patients with RA within Kuwait at large. Finally, we unfortunately were not able to use mobile labs to measure the daily observation of air pollutants in some residential locations close to main hospitals in Kuwait. This forced us to gather air pollution data from fixed monitoring stations, most of which are far away from patients' residences. The reason for this was due to the study's financial costs, which were not funded by any institute or government organisations, but rather all borne by the primary author. Moreover, this prevented us from using VECM to modify a selection of co-variates in the multivariate time series, such as including humidity, wind speed, and wind direction. We did not adjust for them in our study because their data was not accessible per study location or was not genuine.

Ultimately, Johansen's technique gives long-term associations as well as the magnitude of correction needed if the system diverges from the long-term association. If the system converges to a long run equilibrium from short term changes, it also outlines the degree of the error correction term. For the time series SO_2 , NO_2 , O_3 , and DAS28, we employ vector error correction modelling and cointegration to identify functional correlations. They are treated as endogenous variables in our analysis. To test for a unit root, we first utilised the cointegration Regression Dicky Fuller test. In the residuals, the test statistic rules out the existence of a unit root. For the null of no unit root, the test statistic is within the acceptable range. The number of cointegrating vectors was determined using the Johansen trace test. For the rank of the coefficient matrix, those rejected included the null of zero and one or two.

8.7.2 Recommendation and Suggestions

The findings of our research lead to certain recommendations and policy suggestions that can be offered to the Kuwaiti government. Firstly, there are a variety of policy solutions that can aid in the reduction of emissions. One of them is the imposition of pollution taxes. Another strategy to help reduce air pollution levels is to enhance the role of renewable and clean energy consumption, specifically nuclear energy, and of energy efficiency. Besides this, significant variables that control the association between disease activity and air pollution, such as economic output, could be measured by GDP per capita or energy use per capita. Additionally, a spatial method could be used to assess the distance between a residence or workplace and a pollution source.

Suggestions for future research could be the use of novel time series methodology based on machine learning or artificial intelligence with deep learning processes, such as the deep recurrent neural network (DRNN) model, or hybrid deep neural network (HDNN) framework, which is one of the most successful at predicting air pollution (Bhanja and Das, 2021). The Artificial Neural Networks (ANNs) are machine learning approaches, and their basic idea involves constructing a model for mimicking the intelligence of human brain within a machine.

In fact, we highly recommend scholars and researchers consider deep learning for predicting air pollution and any other type of disease activity. With the fast application and development of the sensor technology, air quality prediction is becoming more and more reliant on a variety of sensors and data acquisition equipment to collect large volumes of urban air data regarding pollutants and other factors, such as $PM_{2.5}$, NO_2 , PM_{10} , traffic data, and weather data. Because classic, shallow learning models are limited in their ability to handle large amounts of data, new air quality forecasting methods require data-driven model support (Zhou et al., 2015; Zheng et al., 2013). We highly recommend implementing intelligent computation to predict air pollution and its effect on disease activity scores. In addition, we highly recommend using the Kalman filter to estimate a smoothed trend line through time series consisting of one observation per time point, such as day, month or year. This method depends on a deep learning approach based on structural time series models in combination with the Kalman filter. This is very useful when dealing with missing data to fill long periods of missing values for meteorological observation data (Xie et al., 2021; Hadeed et al., 2020; Afrifa-Yamoah et al., 2020).

Conclusively, the VECM confirms for long and short run equilibrium that there are causal effects from O_3 and NO_2 toward disease activity score for RA patients who are living in a residential area surrounded by the sources of pollutants.

Appendix A

Air Pollution Comparisons

A.1 Air Pollutants' Comparison between Industrial & Residential Stations

Table A.1: Comparison between Industrial Stations using ANOVA test.

	MUT N = 1772	SUB N = 1093	SUK N = 1788	p-Value	N
O_3 (ppm)	0.026 (0.012)	0.019 (0.023)	0.023 (0.011)	<0.001	4595
NO_2 (ppm)	0.024 (0.012)	0.018 (0.018)	0.030 (0.012)	<0.001	4634
NO_x (ppm)	. (.)	0.035 (0.046)	0.052 (0.034)	<0.001	2783
NO (ppm)	0.012 (0.008)	0.016 (0.031)	0.021 (0.025)	<0.001	4432
$W.S.$ (ppm)	2.938 (1.814)	3.354 (2.607)	1.974 (0.727)	<0.001	4498
SO_2 (ppm)	0.004 (0.002)	0.020 (0.032)	0.008 (0.006)	<0.001	4572
CO (ppm)	0.882 (0.345)	0.775 (0.538)	0.358 (0.327)	<0.001	4649
C_6H_6 (ppm)	. (.)	0.001 (0.002)	0.002 (0.002)	<0.001	1339
PM_{10} (g/m ³)	0.259 (0.358)	0.128 (0.189)	0.199 (0.194)	<0.001	2504
$NMHC$ (ppm)	0.473 (0.174)	0.602 (2.488)	0.305 (0.204)	<0.001	3894
Temperature	26.985 (10.260)	23.421 (8.806)	30.323 (9.883)	<0.001	4498
RH	29.364 (17.491)	47.913 (25.320)	33.030 (18.703)	<0.001	4476

Table A.2: Comparison of the Residential Stations using ANOVA test.

	ASA	FAH	JAH	MAN	QUR	RUM	SAA	p-Value
	N = 1750	N = 1817	N = 1818	N = 1767	N = 1189	N = 1810	N = 1726	
O_3 (ppm)	0.022 (0.010)	0.019 (0.009)	0.026 (0.010)	0.024 (0.011)	0.030 (0.015)	0.024 (0.011)	0.026 (0.014)	<0.001
NO_2 (ppm)	0.043 (0.024)	0.051 (0.026)	0.021 (0.010)	0.035 (0.023)	0.041 (0.025)	0.031 (0.020)	0.036 (0.023)	0.000
NO_x (ppm)	0.056 (0.029)	0.078 (0.042)	0.034 (0.036)	.	0.060 (0.041)	0.047 (0.034)	0.048 (0.027)	<0.001
NO (ppm)	0.013 (0.012)	0.025 (0.020)	0.019 (0.054)	0.019 (0.027)	0.019 (0.021)	0.015 (0.018)	0.013 (0.011)	<0.001
SO_2 (ppm)	0.009 (0.005)	0.016 (0.017)	0.005 (0.005)	0.005 (0.004)	0.006 (0.001)	0.009 (0.005)	0.008 (0.008)	0.000
CO (ppm)	0.799 (0.317)	1.297 (0.474)	0.364 (0.406)	0.984 (0.409)	0.826 (0.332)	1.083 (1.679)	0.747 (0.322)	0.000
C_6H_6 (ppm)	0.002 (0.001)	0.002 (0.002)	0.001 (0.001)	.	.	.	0.001 (0.001)	<0.001
PM_{10} (g/m ³)	0.259 (0.316)	0.174 (0.280)	0.169 (0.198)	0.321 (2.380)	.	0.248 (0.229)	0.185 (0.245)	0.002
$NMHC$ (ppm)	0.775 (0.304)	0.627 (0.262)	0.517 (0.761)	0.546 (0.489)	0.512 (0.206)	0.523 (0.211)	0.634 (0.463)	<0.001
W.S.	2.662 (1.290)	3.304 (1.272)	2.614 (1.400)	2.217 (1.022)	3.070 (1.093)	2.312 (0.807)	2.410 (1.200)	<0.001
Temperature	26.984 (9.366)	25.405 (9.288)	28.388 (10.647)	28.282 (9.104)	25.586 (9.212)	27.540 (9.283)	29.617 (9.697)	<0.001
RH	40.179 (23.777)	35.578 (23.866)	28.649 (19.246)	45.755 (16.408)	40.405 (21.757)	58.046 (22.343)	33.962 (21.184)	0.000

Appendix B

Missing Imputation Results

B.1 Figures

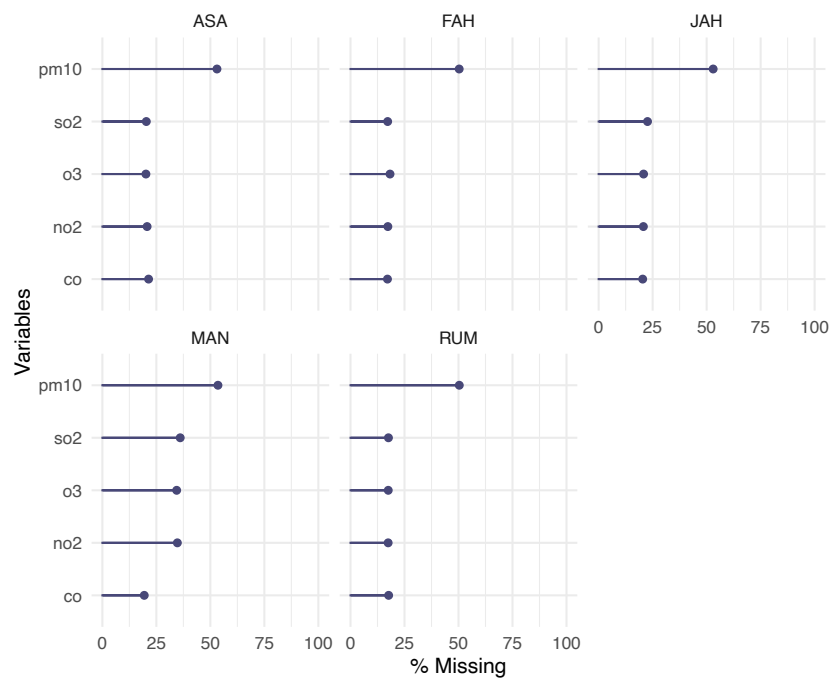


Figure B.1: Missing values for air quality pollutants from 2012 to 2017 per fixed station.

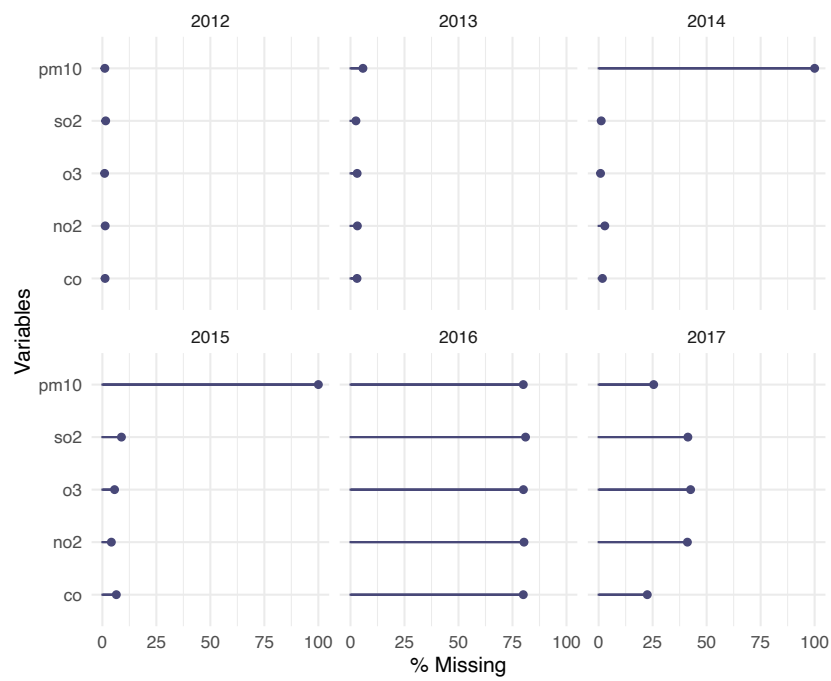


Figure B.2: Missing values for air quality pollutants from 2012 to 2017 per year.

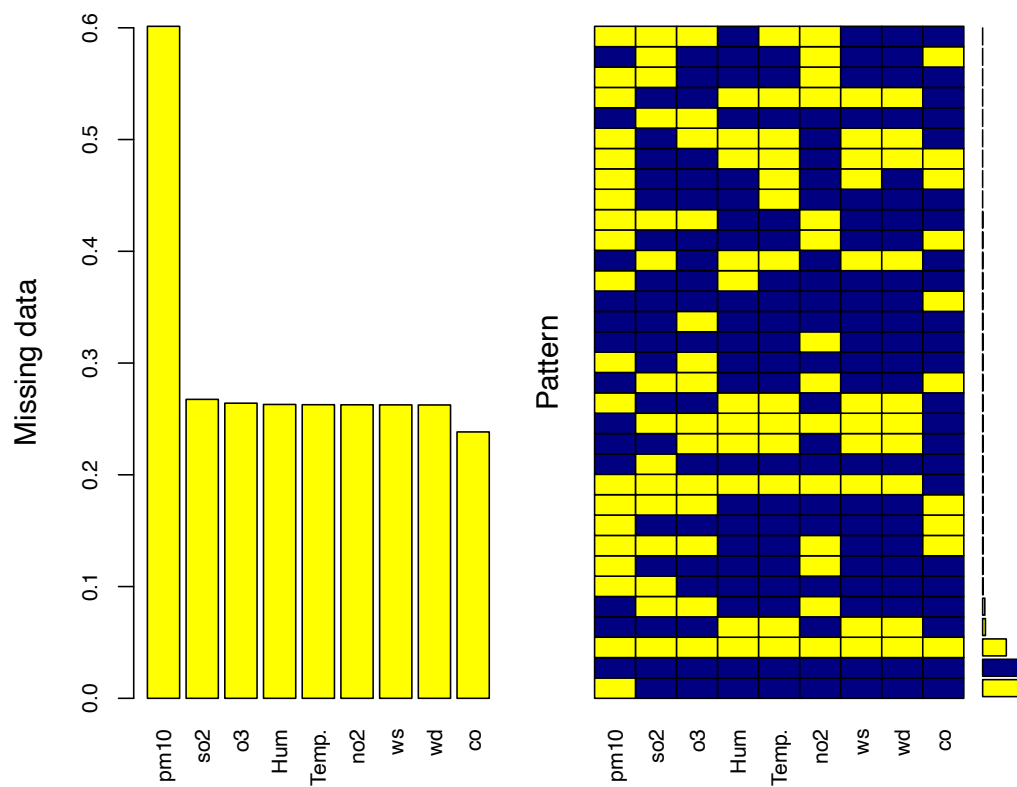


Figure B.3: Missing value patterns for air quality measurements from 2012 to 2017. **Left:** Frequency of missingness in each variable. **Right:** Observed missingness patterns in the data set. The least frequent occurring patterns are located at the top of the plot, with gradually increasing frequency towards the bottom. The y-axis shows the proportion of Non-Missing(Blue) and Missing(Yellow) values.

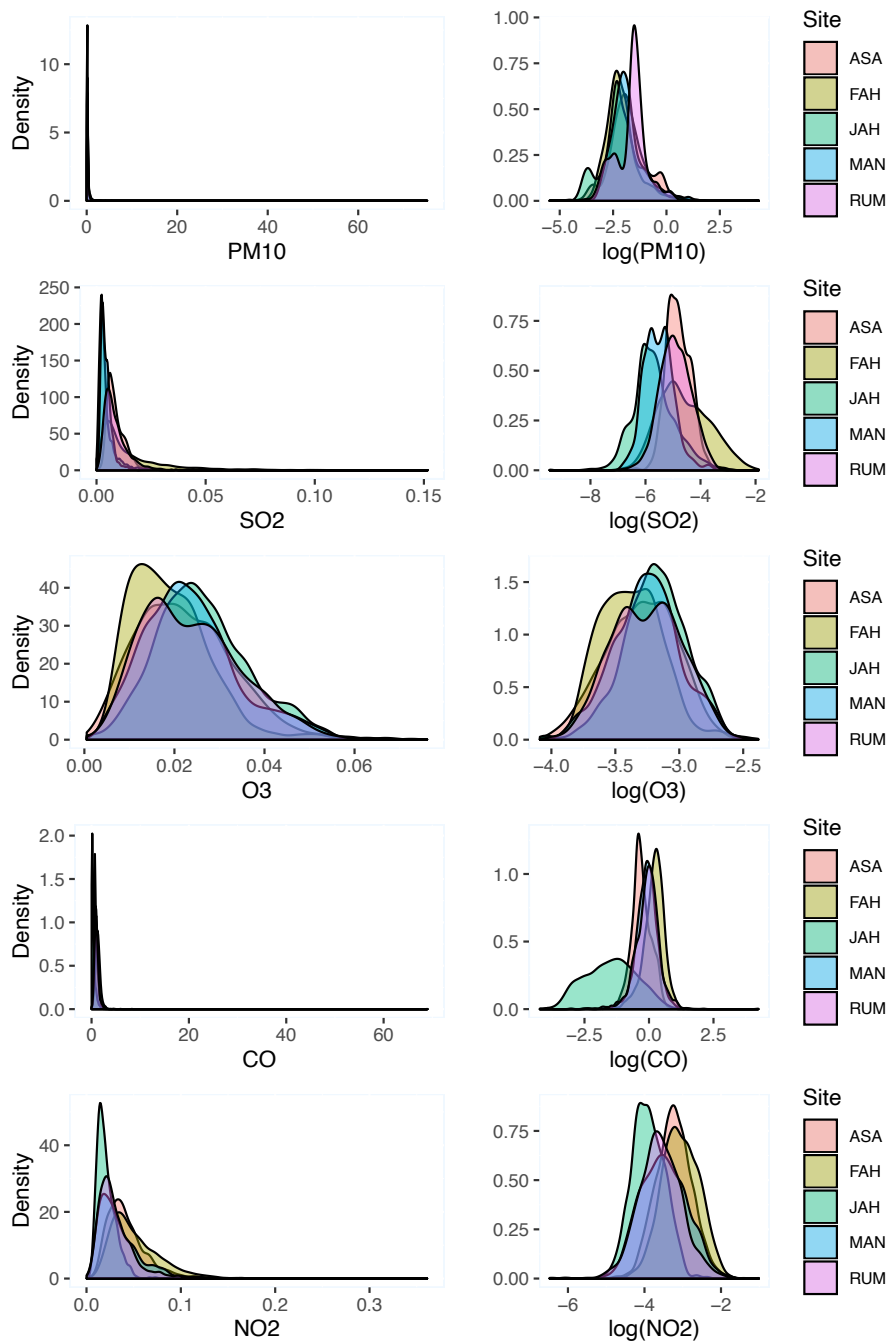


Figure B.4: Distribution analysis for PM_{10} , SO_2 , O_3 , CO , and NO_2 during 2012 to 2017, according to site location in the State of Kuwait. It is very obvious that log transformation fixes the distribution shape for all pollutants. This step is very important—that is, normalizing the skewed data, such that they approximately conform to normality—in order to use them in the imputational calculation for more accurate results (Changyong et al., 2014).



Figure B.5: Mean RMSE and MAE results for the Kuwait Environmental Public Authority (KEPA) data, in order to estimate missing values for SO_2 , NO_2 , CO , and O_3 after eliminating PM_{10} due to a high level of missing values. Results are shown for MCAR (left), MAR (middle), and MNAR (right) data.

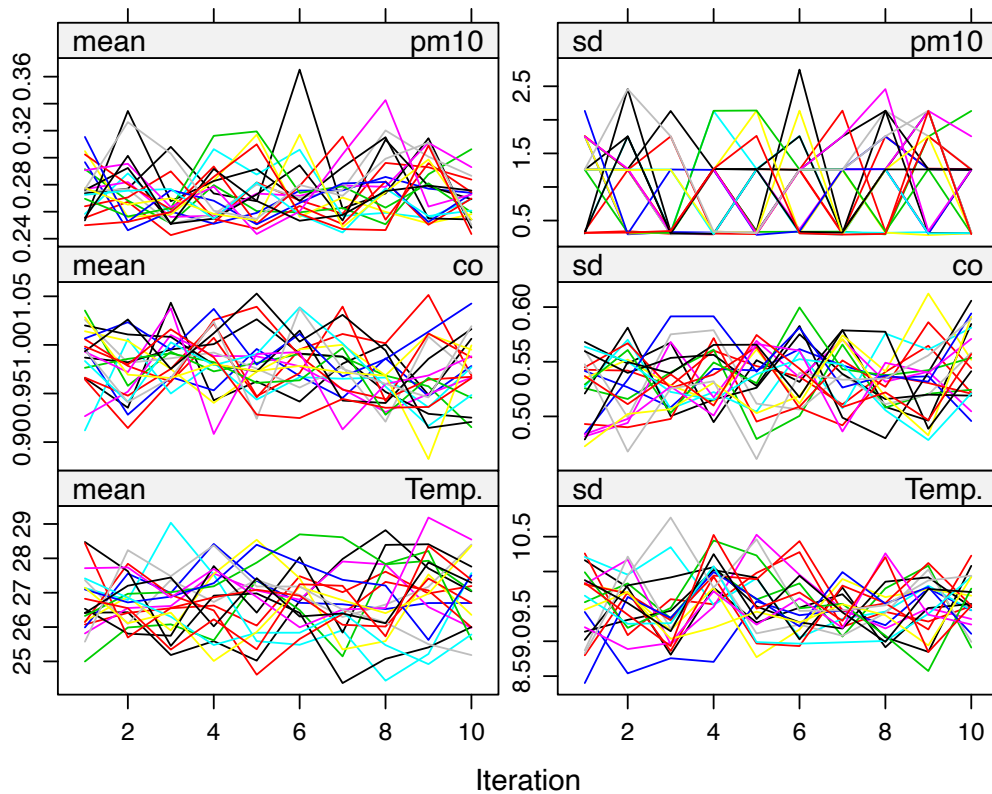


Figure B.6: Inspecting the trace line convergence levels using an iterative Markov Chain–Monte Carlo type of algorithm with respect to the imputed means and standard deviations for PM_{10} , CO , and temperature. These trace plots show the imputed value summaries for all imputed data sets with $m = 20$ after applying 10 iterations, in order to reach to the convergence level of stability.

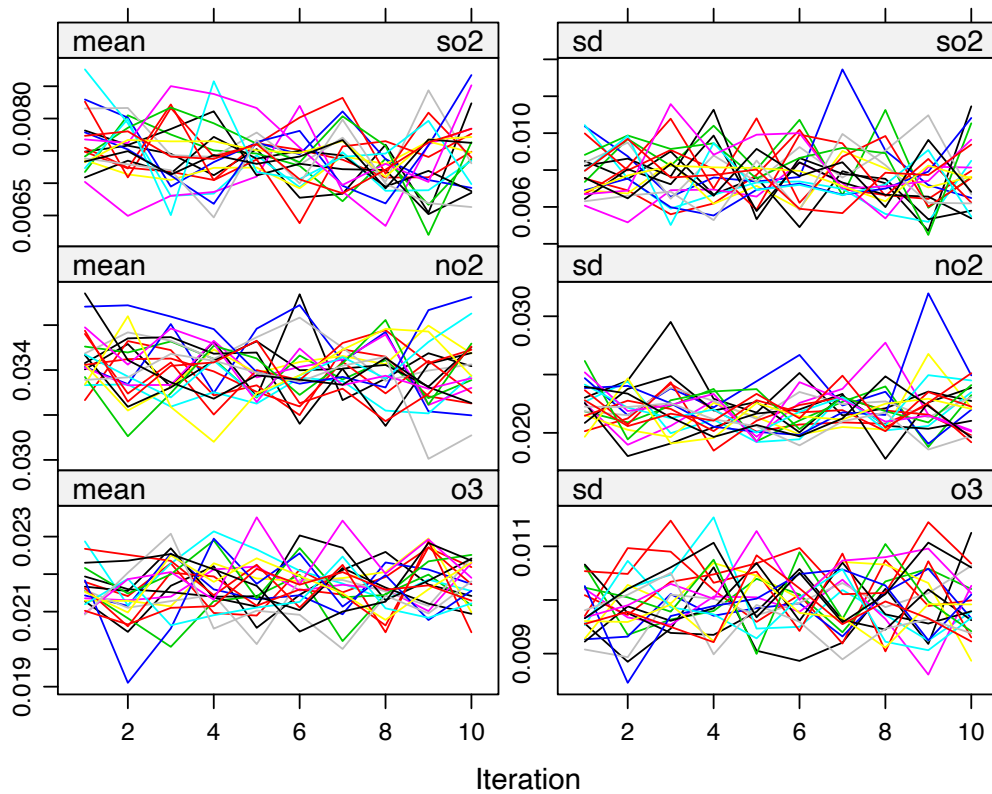


Figure B.7: Inspecting the trace line convergence levels using an iterative Markov Chain–Monte Carlo type of algorithm with respect to the imputed means and standard deviations for SO_2 , NO_2 , and O_3 . These trace plots show the imputed value summaries for all imputed data sets with $m = 20$ after applying 10 iterations, in order to reach to the convergence level of stability. Each colour in the graph represents an imputed data set, where the x-axis represents the number of iterations implemented during the imputational calculation and the y-axis represents the mean (**left**-side) and standard deviation (**right**-side) of the imputed values only.

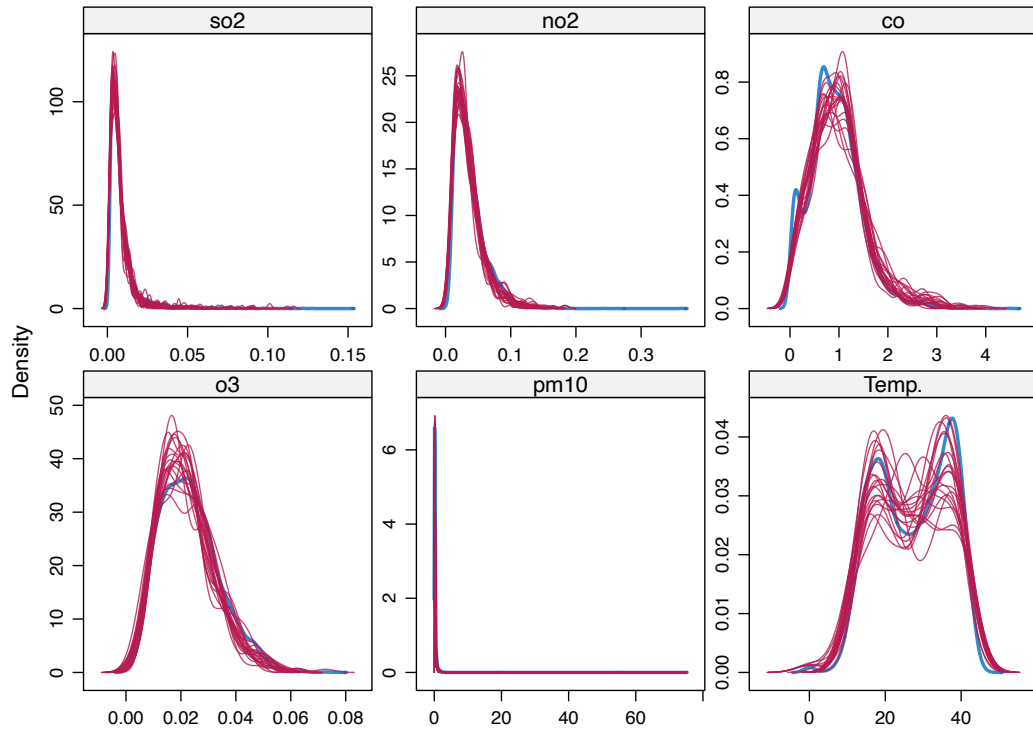
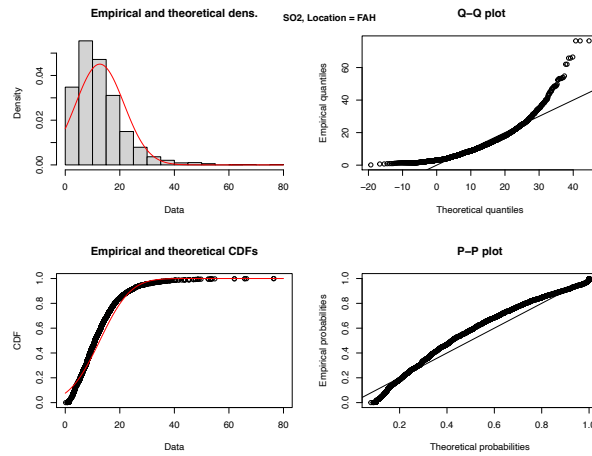


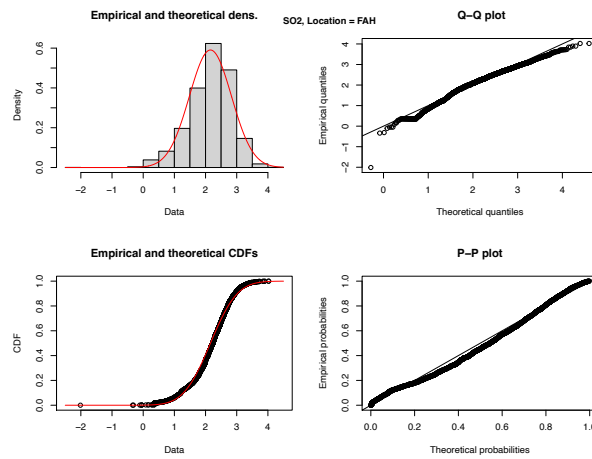
Figure B.8: Density plots with multiple imputations for SO_2 , NO_2 , PM_{10} , CO , and O_3 data. The blue line represents the observed data and the red lines are the density plots of the 20 imputed data sets. As we can see, in all density plots, the red lines almost match the blue line (the observed data), which is an indication of matching between the observed and imputed values.

Appendix C

Normality Assessment

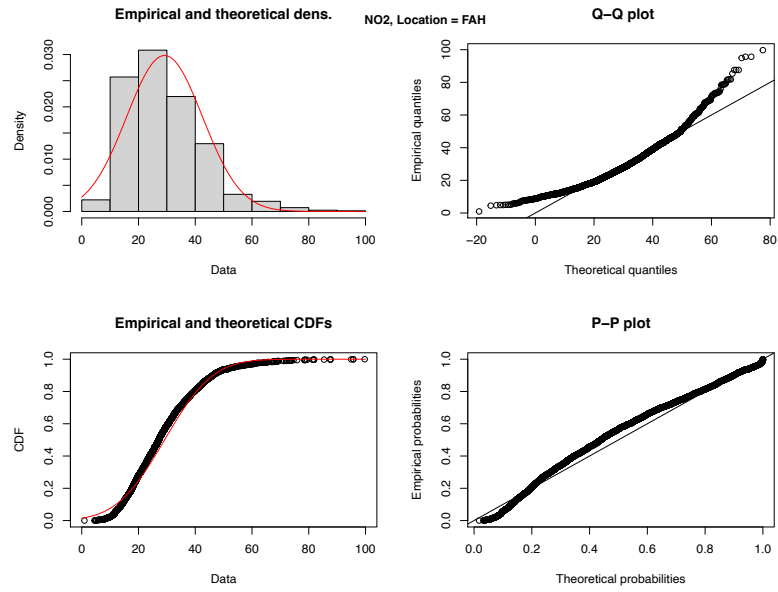


(a) SO_2 measured in FAH location before Box-Cox transformation

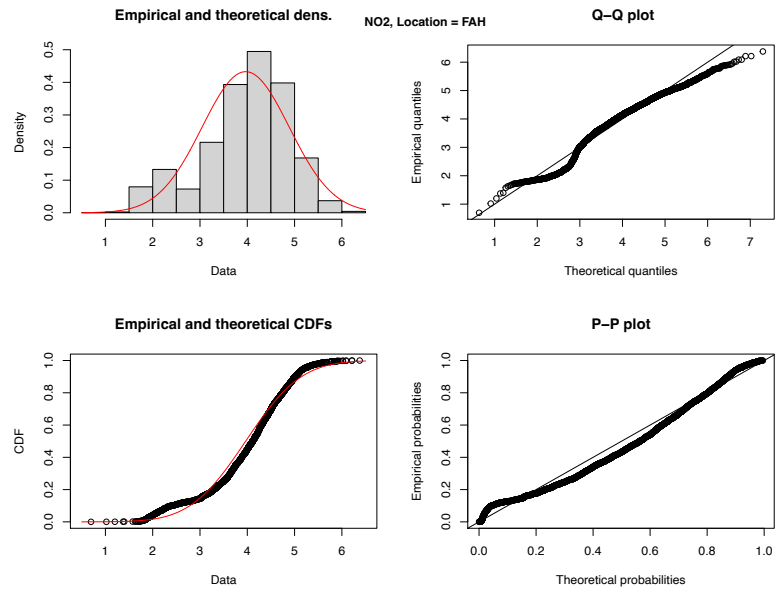


(b) SO_2 measured in FAH location after Box-Cox transformation

Figure C.1: Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for SO_2 before and after the Box-Cox transformation, it is obvious that the Box-Cox transformation enhances the normality performance for SO_2 in FAH location.

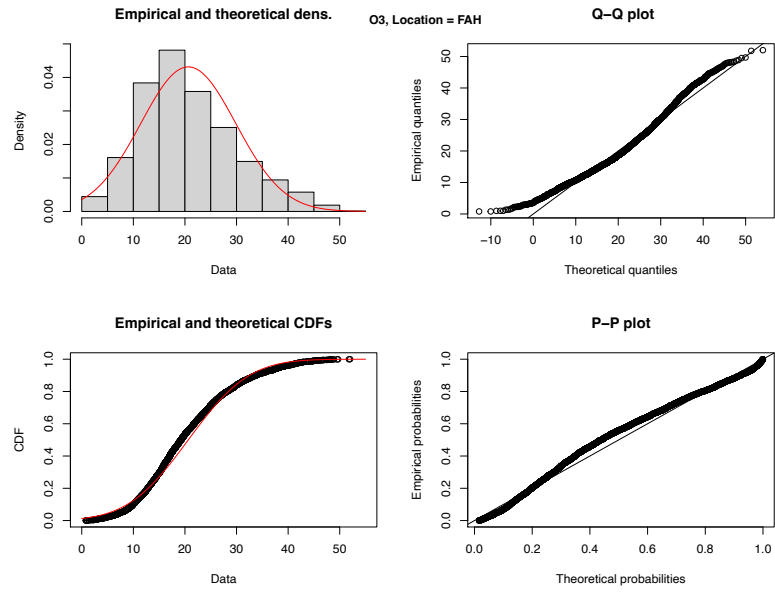


(a) NO_2 measured in FAH location before Box-Cox transformation

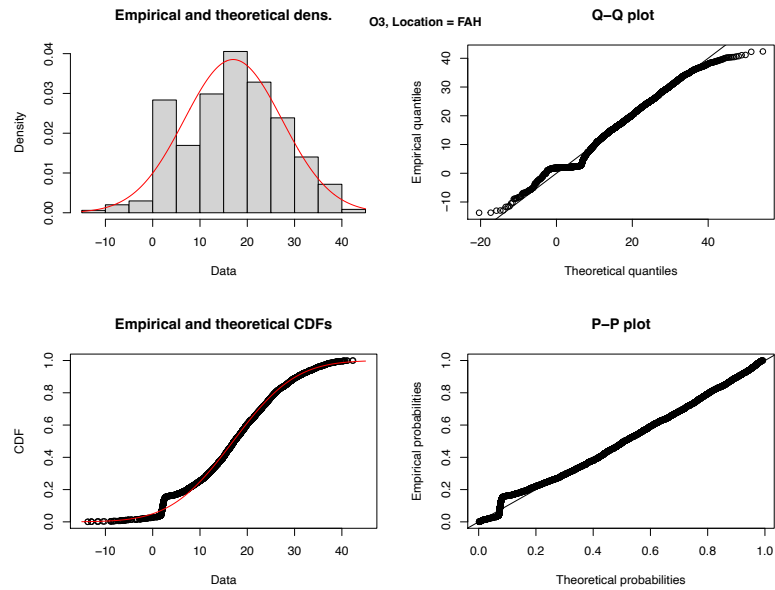


(b) NO_2 measured in FAH location after Box-Cox transformation

Figure C.2: Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for NO_2 before and after the Box-Cox transformation, it is obvious that the Box-Cox transformation enhances the normality performance for NO_2 in FAH location.

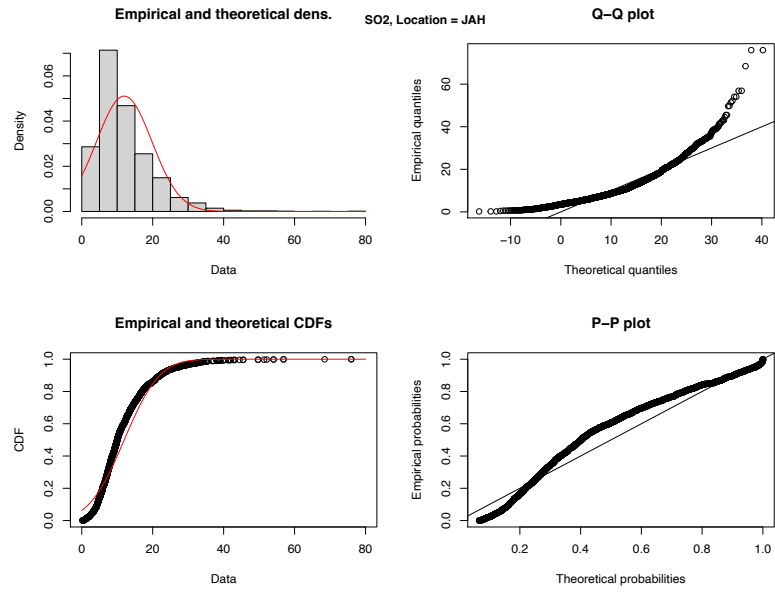


(a) O_3 measured in FAH location before Lambert S transformation

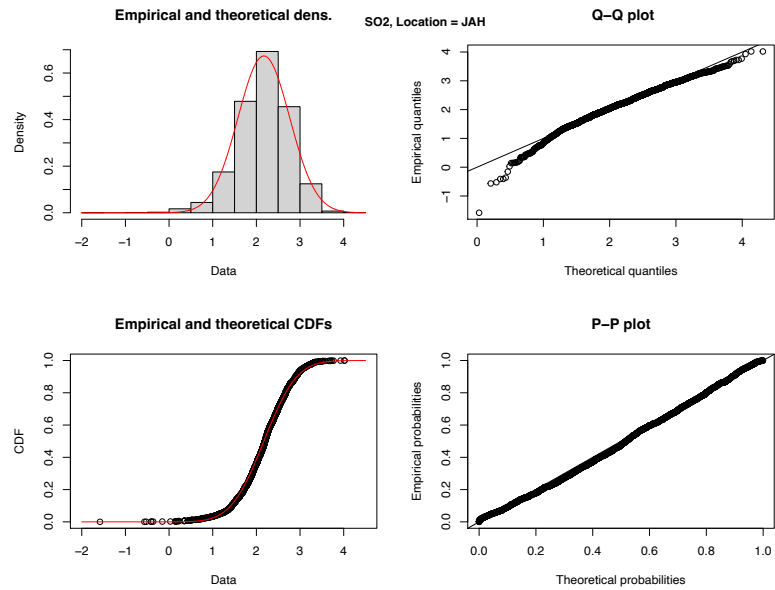


(b) O_3 measured in FAH location after Lambert S transformation

Figure C.3: Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for O_3 before and after the Lambert S transformation, it is obvious that the Lambert S transformation enhances the normality performance for O_3 in FAH location.

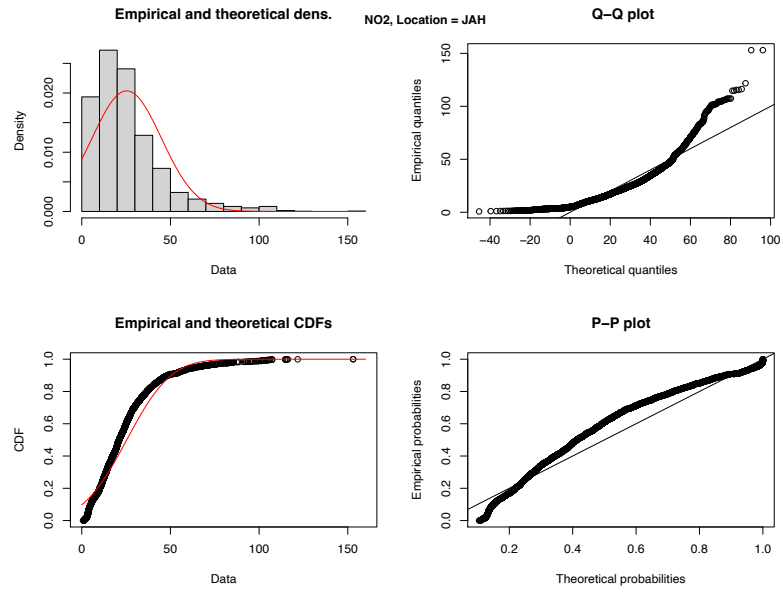


(a) SO_2 measured in JAH location before Yeo-Johnson transformation

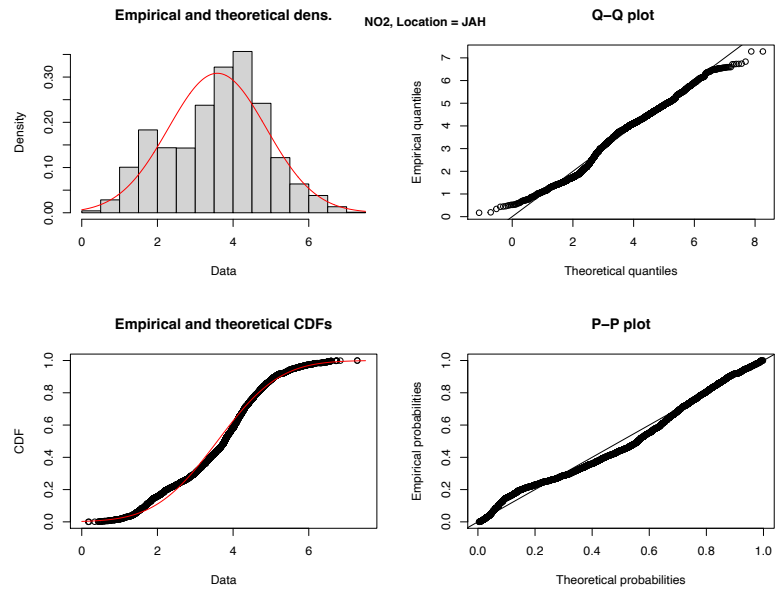


(b) SO_2 measured in JAH location after Yeo-Johnson transformation

Figure C.4: Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for SO_2 before and after the Yeo-Johnson transformation, it is obvious that the Yeo-Johnson transformation enhances the normality performance for SO_2 in JAH location.

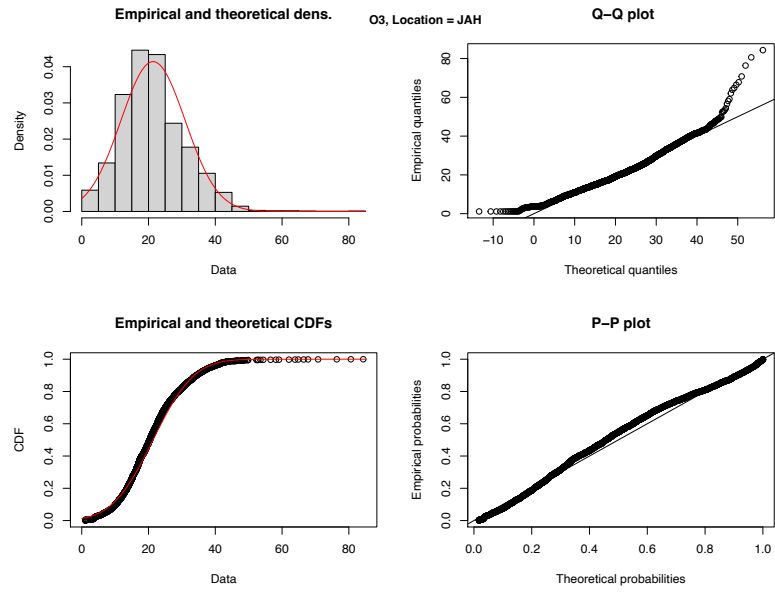


(a) NO_2 measured in JAH location before Yeo-Johnson transformation

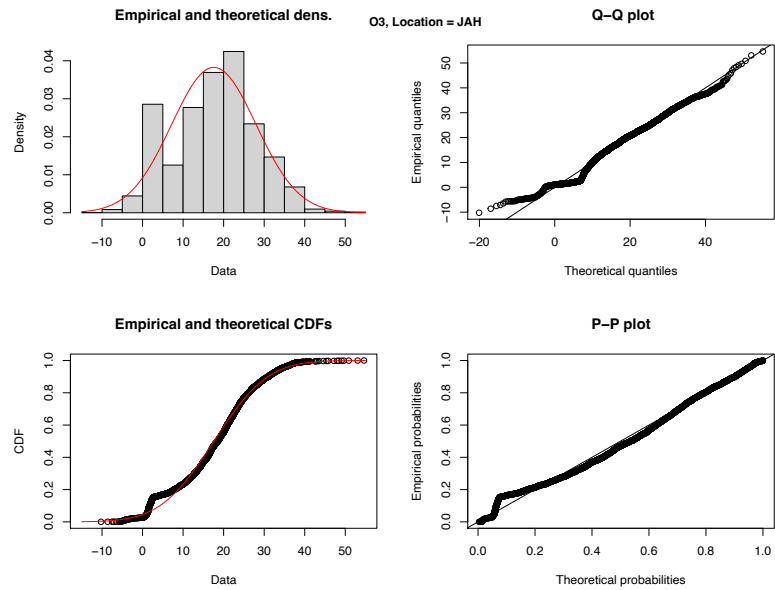


(b) NO_2 measured in JAH location after Yeo-Johnson transformation

Figure C.5: Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for NO_2 before and after the Yeo-Johnson transformation, it is obvious that the Yeo-Johnson transformation enhances the normality performance for NO_2 in JAH location.

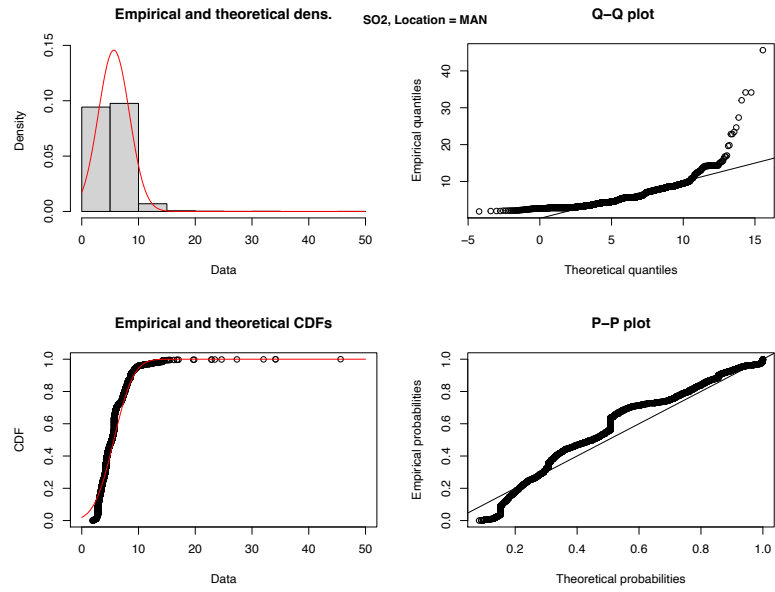


(a) O_3 measured in JAH location before Lambert S transformation

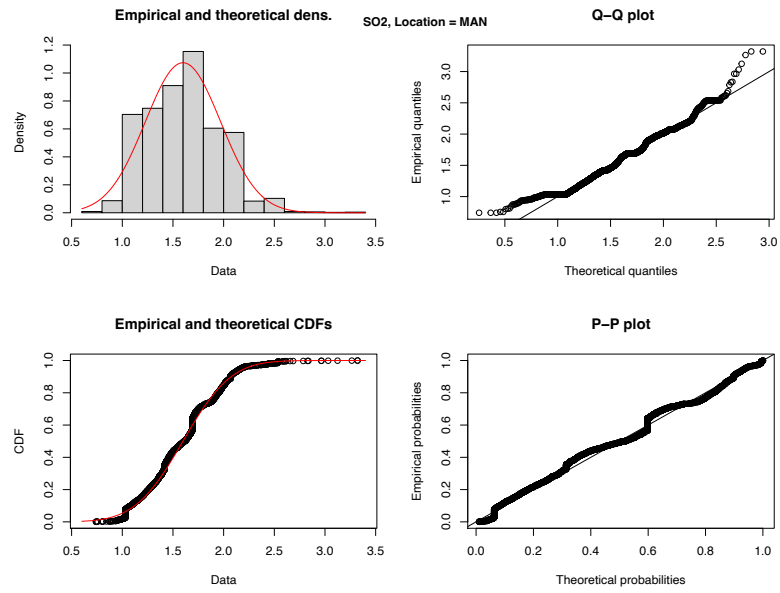


(b) O_3 measured in JAH location after Lambert S transformation

Figure C.6: Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for O_3 before and after the Lambert S transformation, it is obvious that the Lambert S transformation enhances the normality performance for O_3 in JAH location.

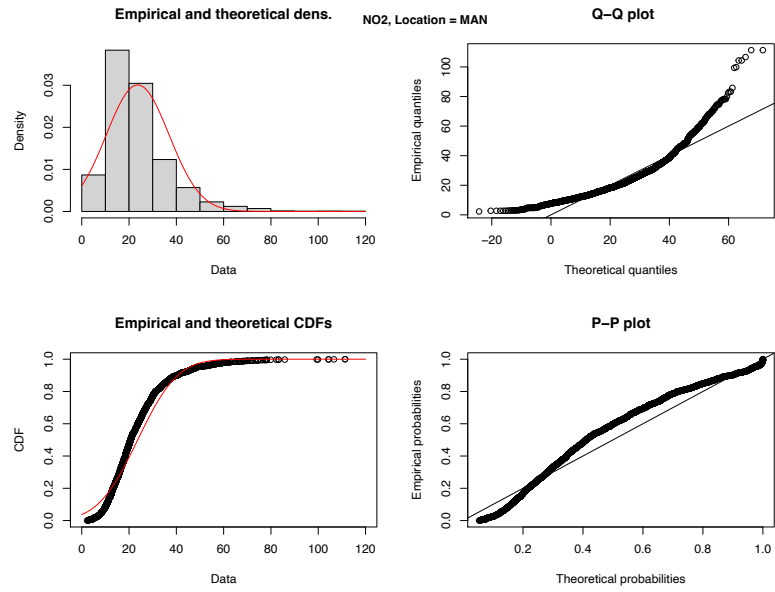


(a) SO_2 measured in MAN location before Log transformation

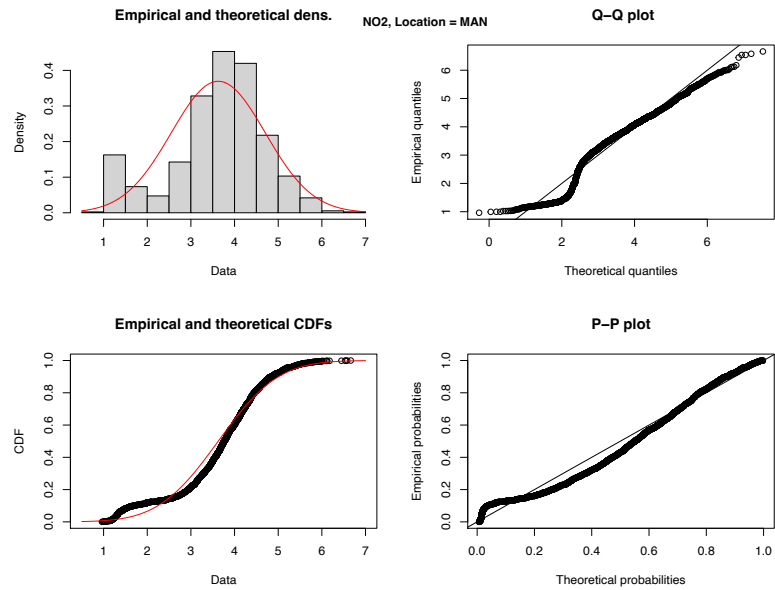


(b) SO_2 measured in MAN location after Log transformation

Figure C.7: Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for SO_2 before and after the Log transformation, it is obvious that the Log transformation enhances the normality performance for SO_2 in MAN location.

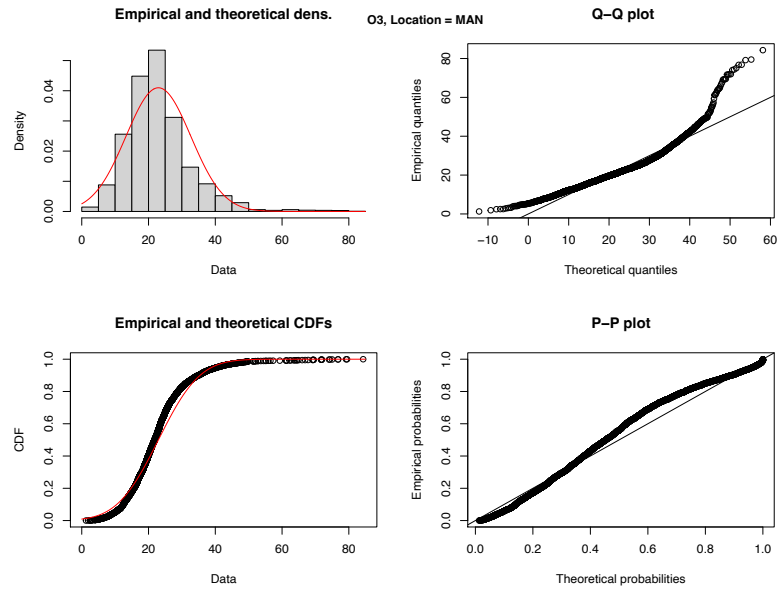


(a) NO_2 measured in MAN location before Yeo-Johnson transformation

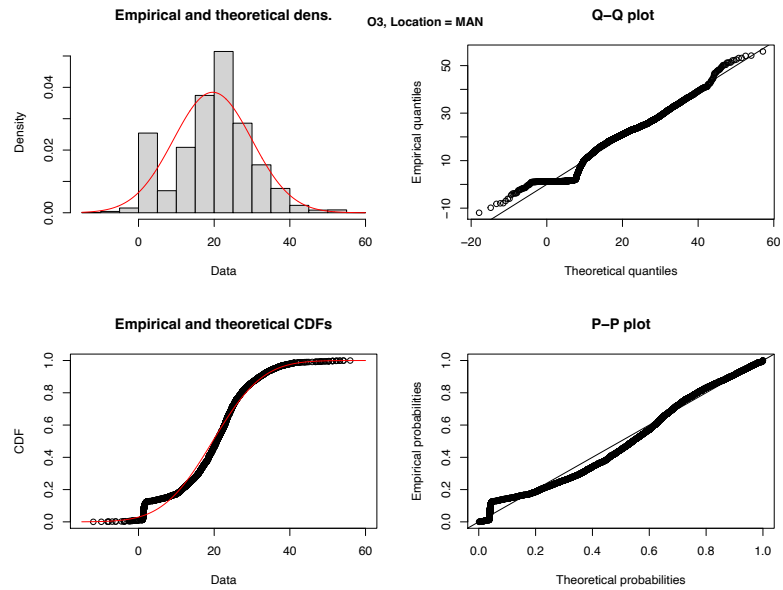


(b) NO_2 measured in MAN location after Yeo-Johnson transformation

Figure C.8: Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for NO_2 before and after the Yeo-Johnson transformation, it is obvious that the Yeo-Johnson transformation enhances the normality performance for NO_2 in MAN location.

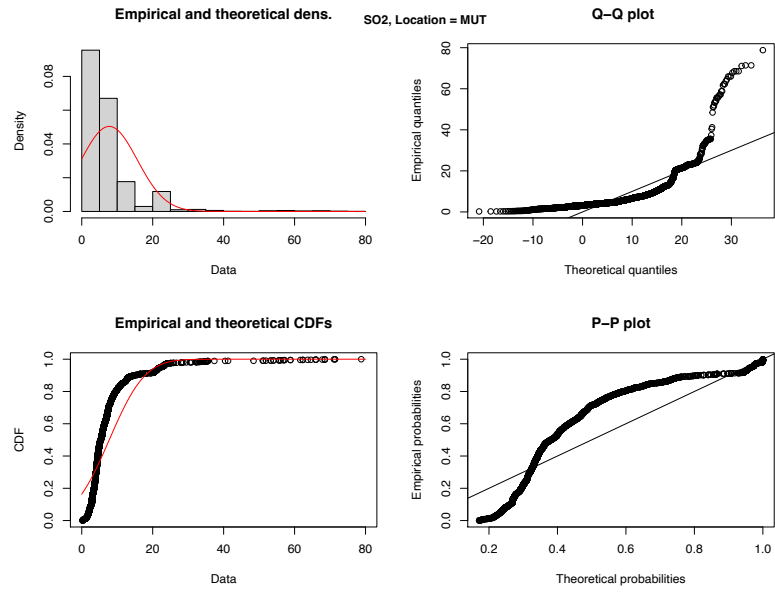


(a) O_3 measured in MAN location before Lambert S transformation

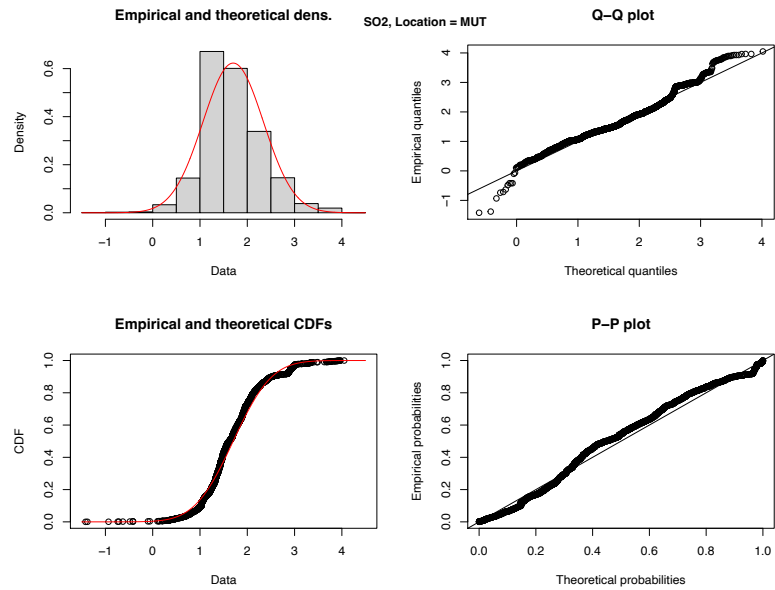


(b) O_3 measured in MAN location after Lambert S transformation

Figure C.9: Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for O_3 before and after the Lambert S transformation, it is obvious that the Lambert S transformation enhances the normality performance for O_3 in MAN location.

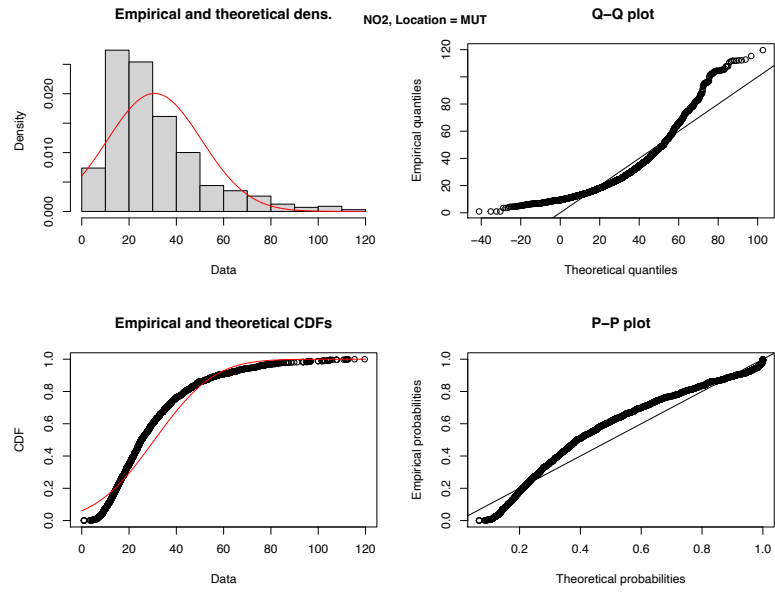


(a) SO_2 measured in MUT location before Yeo-Johnson transformation

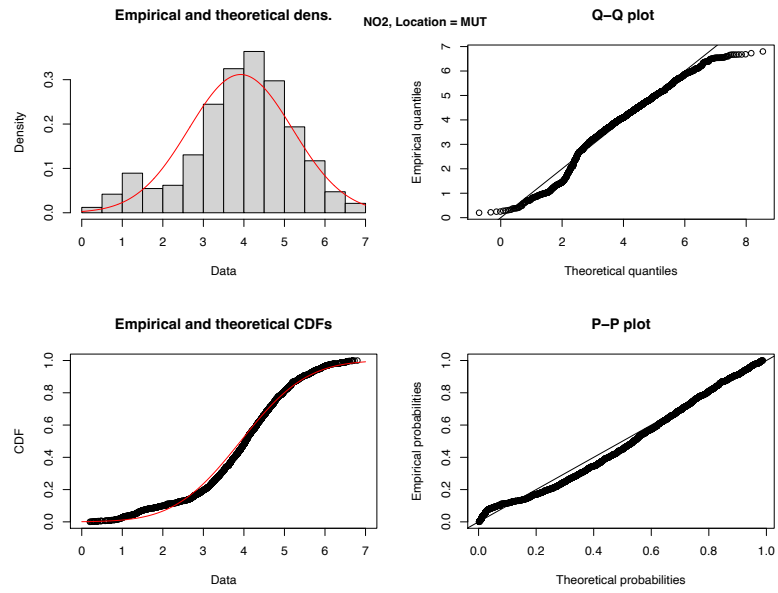


(b) SO_2 measured in MUT location after Yeo-Johnson transformation

Figure C.10: Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for SO_2 before and after the Yeo-Johnson transformation, it is obvious that the Log Transform transformation enhances the normality performance for SO_2 in MUT location.

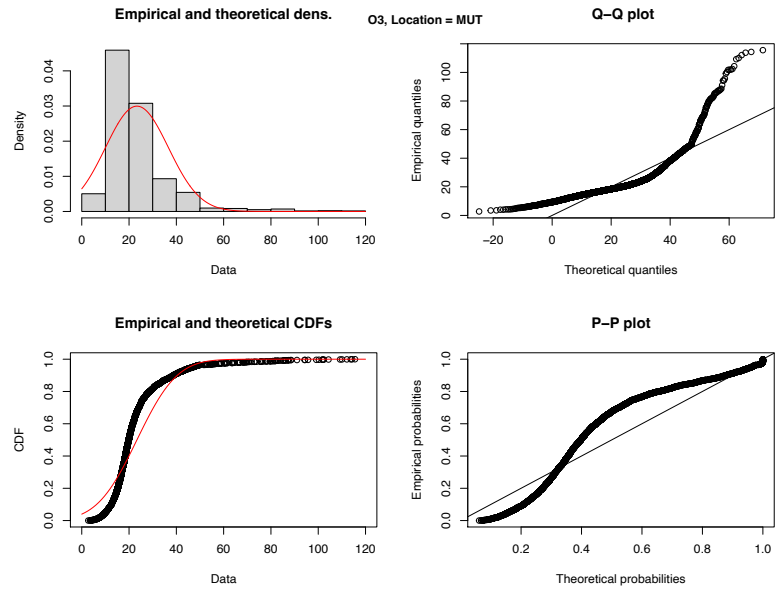


(a) NO_2 measured in MUT location before Log transformation

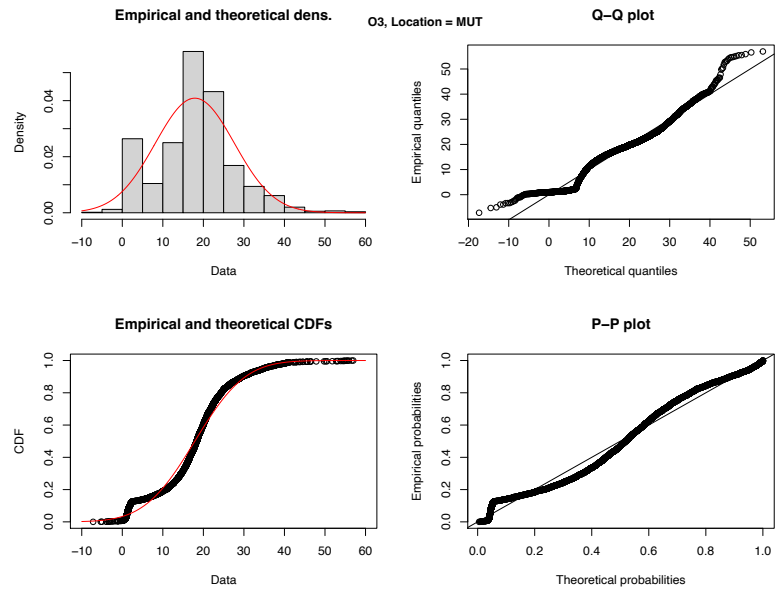


(b) NO_2 measured in MUT location after Log transformation

Figure C.11: Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for NO_2 before and after the Log transformation, it is obvious that the Log transformation enhances the normality performance for NO_2 in MUT location.



(a) O_3 measured in MUT location before Yeo-Johnson transformation



(b) O_3 measured in MUT location after Yeo-Johnson transformation

Figure C.12: Normality plots (Q-Q plot, P-P plot, Histogram and CDF plot) for O_3 before and after the Yeo-Johnson transformation, it is obvious that the Yeo-Johnson transformation enhances the normality performance for O_3 in MUT location.

References

- Abhilash, M., Thakur, A., Gupta, D., and Sreevidya, B. (2018). Time series analysis of air pollution in Bengaluru using ARIMA model. In *Ambient Communications and Computer Systems*, pages 413–426. Springer.
- Achilleos, S., Al-Ozairi, E., Alahmad, B., Garshick, E., Neophytou, A. M., Bouhamra, W., Yassin, M. F., and Koutrakis, P. (2019). Acute effects of air pollution on mortality: A 17-year analysis in Kuwait. *Environment International*, 126:476–483.
- Afoakwah, C., Nghiem, S., Scuffham, P., Huynh, Q., Marwick, T., and Byrnes, J. (2020). Impacts of air pollution on health: evidence from longitudinal cohort data of patients with cardiovascular diseases. *The European Journal of Health Economics*, 21:1025–1038.
- Afrifa-Yamoah, E., Mueller, U. A., Taylor, S., and Fisher, A. (2020). Missing data imputation of high-resolution temporal climate time series data. *Meteorological Applications*, 27(1):e1873.
- Agency, U. E. P. (2015). National ambient air quality standards for ozone; final rule. *Fed. Regist.*, 80(206):65–292.
- Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1):243–247.
- Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265.

References

- Al-Ali, Z., Abdullah, M., Asadalla, N., and Gholoum, M. (2020). A comparative study of remote sensing classification methods for monitoring and assessing desert vegetation using a UAV-based multispectral sensor. *Environmental Monitoring and Assessment*, 192:1–14.
- Al-Awadhi, F. A. and Al-Awadhi, S. A. (2006). Spatial-temporal model for ambient air pollutants in the state of Kuwait. *Environmetrics*, 17(7):739–752.
- Al-Awadhi, J. M. (2014). Measurement of air pollution in Kuwait City using passive samplers. *Atmospheric and Climate Sciences*, 4(02):253.
- Al-Baroud, A., Al-Baroud, F., Al-Sahali, M., and Ettouney, H. (2012). Annual variations of air pollution in Jahra, Kuwait. *GSTF International Journal of Engineering Technology (JET)*, 1:74–79.
- Al-Enezi, E., Al-Dousari, A., and Al-Shammari, F. (2014). Modeling adsorption of inorganic phosphorus on dust fallout in Kuwait Bay. *Journal of Engineering Research*, 2(2):1–14.
- Al-Fadhli, A. A. (2017). Ambient air quality assessment of twelve inhabited areas in the state of Kuwait between years 2011-2014. *International Journal of Chemical Engineering and Applications*, 8(5).
- Al-Fadhli, F. M., Alhajeri, N. S., Aly, A. Z., and Allen, D. T. (2019). The impact of power plant emission variability and fuel switching on the air quality of Kuwait. *Science of the Total Environment*, 672:593–603.
- Al-Hemoud, A., Al-Dousari, A., Al-Shatti, A., Al-Khayat, A., Behbehani, W., and Malak, M. (2018). Health impact assessment associated with exposure to PM_{10} and dust storms in Kuwait. *Atmosphere*, 9(1):6.
- Al-Hemoud, A., Gasana, J., Alajeel, A., Alhamoud, E., Al-Shatti, A., and Al-Khayat, A. (2021). Ambient exposure of O_3 and NO_2 and associated health risk in Kuwait. *Environmental Science and Pollution Research*, 28(12):14917–14926.

References

- Al-Herz, A., Al-Awadhi, A., Saleh, K., Al-Kandari, W., Hasan, E., Ghanem, A., Abutiban, F., Alenizi, A., Hussain, M., Ali, Y., et al. (2016). A comparison of rheumatoid arthritis patients in Kuwait with other populations: results from the krrd registry. *Journal of Advances in Medicine and Medical Research*, pages 1–11.
- Al-Hurban, A., Khader, S., Alsaber, A., and Pan, J. (2021). Air quality assessment in the state of Kuwait during 2012 to 2017. *Atmosphere*, 12(6):678.
- Al Mulla, A., Fanous, N., Seidenberg, A. B., and Rees, V. W. (2015). Secondhand smoke emission levels in water pipe cafes in Doha, Qatar. *Tobacco Control*, 24(e3):e227–e231.
- Al-Mutairi, N., Koushki, P., et al. (2009). Potential contribution of traffic to air pollution in the state of Kuwait. *American Journal of Environmental Sciences*, 5(3):218.
- Al-Salem, S. (2008). An overview of the PM_{10} pollution problem in Fahaheel urban area, Kuwait. *Emirates Journal of Engineering Research*, 13(3):1–9.
- Al-Sarawi, M., Massoud, M., Al-Thoweini, F., and Abdulrassol, A. (2002). Environmental impact assessment of the oil sector complex construction works offshore Al-Shuwaikh coast, Kuwait, II. quality and mercury content of ambient air before construction. *Technology*, 8(1-2):65–77.
- Al-Shayji, K., Lababidi, H., Al-Rushoud, D., and Al-Adwani, H. (2008). Development of a fuzzy air quality performance indicator. *Kuwait Journal of Science and Engineering*, 35:101–126.
- Alahamade, W., Lake, I., Reeves, C. E., and De La Iglesia, B. (2021). Evaluation of multi-variate time series clustering for imputation of air pollution data. *Geoscientific Instrumentation, Methods and Data Systems Discussions*, pages 1–23.
- Albassam, E., Khan, A., and Popov, V. (2009). Management of air quality in the vicinity of congested area in Kuwait. *Environmental Monitoring and Assessment*, 157(1-4):539–555.

References

- Alenezi, R. A. and Al-Anezi, B. S. (2015). An assessment of ambient air quality in two major cities in the state of Kuwait. *International Journal of Engineering & Technology*, 4(2):358.
- Aletaha, D., Nell, V. P., Stamm, T., Uffmann, M., Pflugbeil, S., Machold, K., and Smolen, J. S. (2005). Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. *Arthritis Research & Therapy*, 7(4):R796.
- Aletaha, D., Neogi, T., Silman, A. J., Funovits, J., Felson, D. T., Bingham III, C. O., Birnbaum, N. S., Burmester, G. R., Bykerk, V. P., Cohen, M. D., et al. (2010). 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European league against rheumatism collaborative initiative. *Arthritis & Rheumatism*, 62(9):2569–2581.
- Aletaha, D. and Smolen, J. (2005). The simplified disease activity index (SDAI) and the clinical disease activity index (CDAI): a review of their usefulness and validity in rheumatoid arthritis. *Clinical and Experimental Rheumatology*, 23(5):S100.
- Alexander, C. (2001). *Market models: A Guide to Financial Data Analysis*. John Wiley Sons Ltd., Chichester.
- Allen, R. J. and DeGaetano, A. T. (2001). Estimating missing daily temperature extremes using an optimized regression approach. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 21(11):1305–1319.
- Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods & Research*, 28(3):301–309.
- Allison, P. D. (2001). *Missing Data*. Sage publications.
- Alsaber, A., Al-Herz, A., Pan, J., AL-Sultan, A. T., Mishra, D., and Group, K. (2021a). Handling missing data in a rheumatoid arthritis registry using random forest approach. *International Journal of Rheumatic Diseases*, 24(10):1282–1293.

References

- Alsaber, A., Pan, J., Al-Herz, A., Alkandary, D. S., Al-Hurban, A., Setiya, P., Group, K., et al. (2020). Influence of ambient air pollution on rheumatoid arthritis disease activity score index. *International Journal of Environmental Research and Public Health*, 17(2):416.
- Alsaber, A. R., Pan, J., and Al-Hurban, A. (2021b). Handling complex missing data using random forest approach for an air quality monitoring dataset: a case study of Kuwait environmental data (2012 to 2018). *International Journal of Environmental Research and Public Health*, 18(3):1333.
- Amoatey, P., Omidvarborna, H., and Baawain, M. (2018). The modeling and health risk assessment of $PM_{2.5}$ from Tema oil refinery. *Human and Ecological Risk Assessment: An International Journal*, 24(5):1181–1196.
- Anenberg, S. C., Henze, D. K., Tinney, V., Kinney, P. L., Raich, W., Fann, N., Malley, C. S., Roman, H., Lamsal, L., Duncan, B., et al. (2018). Estimates of the global burden of ambient $PM_{2.5}$, O_3 , and NO_2 on asthma incidence and emergency room visits. *Environmental Health Perspectives*, 126(10):107004.
- Argyropoulos, C. D., Abraham, M., Hassan, H., Ashraf, A., Fthenou, E., Sadoun, E., and Kakosimos, K. (2016). Modeling of PM_{10} and $PM_{2.5}$ building infiltration during a dust event in Doha, Qatar. In *Proceedings of 2nd International Conference on Atmospheric Dust–DUST2016. Castellaneta Marina-Taranto, Italy*.
- Arize, A. C. (2017). A convenient method for the estimation of ARDL parameters and test statistics: USA trade balance and real effective exchange rate relation. *International Review of Economics & Finance*, 50:75–84.
- Asari, F., Baharuddin, N. S., Jusoh, N., Mohamad, Z., Shamsudin, N., and Jusoff, K. (2011). A vector error correction model (VECM) approach in explaining the relationship between interest rate and inflation towards exchange rate volatility in Malaysia. *World Applied Sciences Journal*, 12(3):49–56.

References

- Ayres, J. G., Maynard, R. L., and Richards, R. J. (2006). *Air Pollution and Health*, volume 3. World Scientific.
- Azlina, A., Law, S. H., and Mustapha, N. H. N. (2014). Dynamic linkages among transport energy consumption, income and CO_2 emission in Malaysia. *Energy Policy*, 73:598–606.
- Bagheri, H., Tapak, L., Karami, M., Amiri, B., and Cherghi, Z. (2019). Epidemiological features of human brucellosis in Iran (2011-2018) and prediction of brucellosis with data-mining models. *Journal of Research in Health Sciences*, 19(4):e00462.
- Baraldi, A. N. and Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1):5–37.
- Barkley, Z. R., Lauvaux, T., Davis, K. J., Deng, A., Miles, N. L., Richardson, S. J., Cao, Y., Sweeney, C., Karion, A., Smith, M., et al. (2017). Quantifying methane emissions from natural gas production in north-eastern Pennsylvania. *Atmospheric Chemistry and Physics (Online)*, 17(22):13941–13966.
- Barnes, J. H., Chatterton, T. J., and Longhurst, J. W. (2019). Emissions vs exposure: Increasing injustice from road traffic-related air pollution in the United Kingdom. *Transportation Research Part D: Transport and Environment*, 73:56–66.
- Bashir, M. F., Benghouli, M., Numan, U., Shakoor, A., Komal, B., Bashir, M. A., Bashir, M., Tan, D., et al. (2020a). Environmental pollution and COVID-19 outbreak: insights from Germany. *Air Quality, Atmosphere & Health*, 13(11):1385–1394.
- Bashir, M. F., Ma, B., Komal, B., Bashir, M. A., Tan, D., Bashir, M., et al. (2020b). Correlation between climate indicators and COVID-19 pandemic in New York, USA. *Science of the Total Environment*, 728:138835.
- Becker, R., Enders, W., and Hurn, S. (2004). A general test for time dependence in parameters. *Journal of Applied Econometrics*, 19(7):899–906.

References

- Beeson, W. L., Abbey, D. E., and Knutsen, S. F. (1998). Long-term concentrations of ambient air pollutants and incident lung cancer in California adults: results from the AHSMOG study. Adventist Health Study on Smog. *Environmental Health Perspectives*, 106(12):813–823.
- Bell, J. N. B. and Treshow, M. (2002). *Air Pollution and Plant Life*. John Wiley & Sons.
- Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T., Norton, J. P., Perrin, C., et al. (2013). Characterising performance of environmental models. *Environmental Modelling & Software*, 40:1–20.
- Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., and Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in Brief*, 29:105340.
- Bernatsky, S., Smargiassi, A., Barnabe, C., Svenson, L. W., Brand, A., Martin, R. V., Hudson, M., Clarke, A. E., Fortin, P. R., van Donkelaar, A., et al. (2016). Fine particulate air pollution and systemic autoimmune rheumatic disease in two Canadian provinces. *Environmental Research*, 146:85–91.
- Bertsimas, D., Pawlowski, C., and Zhuo, Y. D. (2017). From predictive methods to missing data imputation: an optimization approach. *The Journal of Machine Learning Research*, 18(1):7133–7171.
- Bhanja, S. and Das, A. (2021). A hybrid deep learning model for air quality time series prediction. *Indonesian Journal of Electrical Engineering and Computer Science*, 22(3):1611–1618.
- Birmili, W., R uckerl, R., Hoffmann, B., Weinmayr, G., Schins, R., Kuhlbusch, T., Vogel, A., Weber, K., Franck, U., Cyrys, J., et al. (2014). Ultrafeine aerosolpartikel in der au enluft: Perspektiven zur aufkl rung ihrer gesundheitseffekte. *Gefahrst Reinhalt Luft*, 74:429–500.

References

- Blough, S. R. (1992). The relationship between power and level for generic unit root tests in finite samples. *Journal of Applied Econometrics*, 7(3):295–308.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.
- Bouhamra, W. and Abdul-Wahab, S. (1999). Description of outdoor air quality in a typical residential area in Kuwait. *Environmental Pollution*, 105(2):221–229.
- Box, G. E. and Jenkins, G. M. (1968). Some recent advances in forecasting and control. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 17(2):91–109.
- Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.
- Branden, K. V. and Verboven, S. (2009). Robust data imputation. *Computational Biology and Chemistry*, 33(1):7–13.
- Brauer, M., Amann, M., Burnett, R. T., Cohen, A., Dentener, F., Ezzati, M., Henderson, S. B., Krzyzanowski, M., Martin, R. V., Van Dingenen, R., et al. (2012). Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution. *Environmental Science & Technology*, 46(2):652–660.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- Briz-Redón, Á. and Serrano-Aroca, Á. (2020). A spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain. *Science of the Total Environment*, 728:138811.
- Brook, R. D., Newby, D. E., and Rajagopalan, S. (2017). The global threat of outdoor ambient air pollution to cardiovascular health: time for intervention. *JAMA Cardiology*, 2(4):353–354.
- Brooks, C. (2019). *Introductory Econometrics for Finance*. Cambridge University Press.

References

- Brooks, C. et al. (2008). RATS handbook to accompany introductory econometrics for finance. *Cambridge Books*.
- Brusseau, M., Ramirez-Andreotta, M., Pepper, I., and Maximillian, J. (2019). Environmental impacts on human health and well-being. In *Environmental and Pollution Science*, pages 477–499. Elsevier.
- Butland, B. K., Samoli, E., Atkinson, R. W., Barratt, B., and Katsouyanni, K. (2019). Measurement error in a multi-level analysis of air pollution and health: a simulation study. *Environmental Health*, 18(1):1–10.
- Buuren, S. v. and Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 85:1–68.
- Capan, M., Hoover, S., Jackson, E. V., Paul, D., and Locke, R. (2016). Time series analysis for forecasting hospital census: Application to the neonatal intensive care unit. *Applied Clinical Informatics*, 7(2):275.
- Carroll, R., Chen, R., Li, T., Newton, H., Schmiediche, H., Wang, N., and George, E. (1997). Trends in ozone exposure in Harris county, Texas. *Journal of the American Statistical Association*, 92:392–415.
- Carslaw, D. C. and Ropkins, K. (2012). Openair—an R package for air quality data analysis. *Environmental Modelling & Software*, 27:52–61.
- Casanova, L. M., Jeon, S., Rutala, W. A., Weber, D. J., and Sobsey, M. D. (2010). Effects of air temperature and relative humidity on coronavirus survival on surfaces. *Applied and Environmental Microbiology*, 76(9):2712.
- Casdagli, M. (1992). Chaos and deterministic versus stochastic non-linear modelling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(2):303–328.
- Chadsuthi, S., Modchang, C., Lenbury, Y., Iamsirithaworn, S., and Triampo, W. (2012). Modeling seasonal leptospirosis transmission and its association with rainfall and tem-

References

- perature in Thailand using time-series and ARIMAX analyses. *Asian Pacific Journal of Tropical Medicine*, 5(7):539–546.
- Chang, K.-H., Hsu, C.-C., Muo, C.-H., Hsu, C. Y., Liu, H.-C., Kao, C.-H., Chen, C.-Y., Chang, M.-Y., and Hsu, Y.-C. (2016). Air pollution exposure increases the risk of rheumatoid arthritis: a longitudinal and nationwide study. *Environment International*, 94:495–499.
- Changyong, F., Hongyue, W., Naiji, L., Tian, C., Hua, H., Ying, L., et al. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, 26(2):105.
- Chaudhari, K., Rizvi, S., and Syed, B. A. (2016). Rheumatoid arthritis: current and future trends. *Nature Reviews Drug Discovery*, 15(5):305–306.
- Chen, R., Wang, X., Meng, X., Hua, J., Zhou, Z., Chen, B., and Kan, H. (2013). Communicating air pollution-related health risks to the public: An application of the air quality health index in Shanghai, China. *Environment International*, 51:168–173.
- Cheng, W.-L., Chen, Y.-S., Zhang, J., Lyons, T., Pai, J.-L., and Chang, S.-H. (2007). Comparison of the revised air quality index with the PSI and AQI indices. *Science of the Total Environment*, 382(2-3):191–198.
- Chevillon, G. and Hendry, D. F. (2005). Non-parametric direct multi-step estimation for forecasting economic processes. *International Journal of Forecasting*, 21(2):201–218.
- Choe, J.-Y., Bae, J., Lee, H., Bae, S.-C., and Kim, S.-K. (2013). Relation of rheumatoid factor and anti-cyclic citrullinated peptide antibody with disease activity in rheumatoid arthritis: cross-sectional study. *Rheumatology International*, 33(9):2373–2379.
- Choi, I., Mok, K., and Tam, S. (2002). Solving harmonic sea-level model with Kalman filter: a Macau case study. In *Carbonate Beaches 2000*, pages 38–52.
- Clavel, J., Merceron, G., and Escarguel, G. (2014). Missing data estimation in morphometrics: how much is too much? *Systematic Biology*, 63(2):203–218.

References

- Cochrane, J. H. (1991). A critique of the application of unit root tests. *Journal of Economic Dynamics and Control*, 15(2):275–284.
- Cochrane, J. H. (2005). Time series for macroeconomics and finance. *Manuscript, University of Chicago*, pages 1–136.
- Cook, R. D. and Weisberg, S. (1999). Graphs in statistical analysis: Is the medium the message? *The American Statistician*, 53(1):29–37.
- Costa, S., Ferreira, J., Silveira, C., Costa, C., Lopes, D., Relvas, H., Borrego, C., Roebeling, P., Miranda, A. I., and Paulo Teixeira, J. (2014). Integrating health on air quality assessment—review report on health risks of two major european outdoor air pollutants: Pm and NO_2 . *Journal of Toxicology and Environmental Health, Part B*, 17(6):307–340.
- Covic, T., Tyson, G., Spencer, D., and Howe, G. (2006). Depression in rheumatoid arthritis patients: demographic, clinical, and psychological predictors. *Journal of Psychosomatic Research*, 60(5):469–476.
- Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.
- Cox, D. R., Gudmundsson, G., Lindgren, G., Bondesson, L., Harsaae, E., Laake, P., Juselius, K., and Lauritzen, S. L. (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, 1981:93–115.
- Cui, Y., Zhang, Z.-F., Froines, J., Zhao, J., Wang, H., Yu, S.-Z., and Detels, R. (2003). Air pollution and case fatality of SARS in the People’s Republic of China: an ecologic study. *Environmental Health*, 2(1):1–5.
- Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). Random forests. In *Ensemble Machine Learning*, pages 157–175. Springer.

References

- Dales, R., Blanco-Vidal, C., Romero-Meza, R., Schoen, S., Lukina, A., and Cakmak, S. (2021). The association between air pollution and COVID-19 related mortality in Santiago, Chile: A daily time series analysis. *Environmental Research*, page 111284.
- Davies, N., Triggs, C., and Newbold, P. (1977). Significance levels of the Box-pierce portmanteau statistic in finite samples. *Biometrika*, 64(3):517–522.
- De Roos, A. J., Koehoorn, M., Tamburic, L., Davies, H. W., and Brauer, M. (2014). Proximity to traffic, ambient air pollution, and community noise in relation to incident rheumatoid arthritis. *Environmental Health Perspectives*, 122(10):1075–1080.
- Deleawe, S., Kuszniir, J., Lamb, B., and Cook, D. J. (2010). Predicting air quality in smart environments. *Journal of Ambient Intelligence and Smart Environments*, 2(2):145–154.
- Di Zio, M., Guarnera, U., and Luzi, O. (2007). Imputation through finite Gaussian mixture models. *Computational Statistics & Data Analysis*, 51(11):5305–5316.
- Díaz-Robles, L. A., Ortega, J. C., Fu, J. S., Reed, G. D., Chow, J. C., Watson, J. G., and Moncada-Herrera, J. A. (2008). A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. *Atmospheric Environment*, 42(35):8331–8340.
- Dickey, D. A. (1976). *Estimation and Hypothesis Testing in Nonstationary Time Series*. Iowa State University.
- Dickey, D. A. and Fuller, W. A. (1979a). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431.
- Dickey, D. A. and Fuller, W. A. (1979b). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431.

References

- Dickey, D. A. and Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica: Journal of the Econometric Society*, (49):1057–1072.
- Distante, C., Piscitelli, P., and Miani, A. (2020). COVID-19 outbreak progression in italian regions: approaching the peak by the end of March in northern Italy and first week of April in southern Italy. *International Journal of Environmental Research and Public Health*, 17(9):3025.
- Doan, M. and East, C. (1977). A proposed air quality index for urban areas. *Water, Air, and Soil Pollution*, 8(4):441–451.
- Dobson, A. J. and Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC.
- Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris Jr, B. G., and Speizer, F. E. (1993). An association between air pollution and mortality in six US cities. *New England Journal of Medicine*, 329(24):1753–1759.
- Domingo, J. L., Marquès, M., and Rovira, J. (2020). Influence of airborne transmission of SARS-CoV-2 on COVID-19 pandemic. a review. *Environmental Research*, page 109861.
- Domingo, J. L. and Rovira, J. (2020). Effects of air pollutants on the transmission and severity of respiratory viral infections. *Environmental Research*, 187:109650.
- Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L., and Samet, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA*, 295(10):1127–1134.
- Dong, M., Yang, D., Kuang, Y., He, D., Erdal, S., and Kenski, D. (2009). $PM_{2.5}$ concentration prediction using hidden semi-Markov model-based times series data mining. *Expert Systems with Applications*, 36(5):9046–9055.

References

- Donnelly, A., Misstear, B., and Broderick, B. (2015). Real time air quality forecasting using integrated parametric and non-parametric regression techniques. *Atmospheric Environment*, 103:53–65.
- Du, W., Li, X., Chen, Y., and Shen, G. (2018). Household air pollution and personal exposure to air pollutants in rural China—a review. *Environmental Pollution*, 237:625–638.
- Earnest, A., Chen, M. I., Ng, D., and Sin, L. Y. (2005). Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. *BMC Health Services Research*, 5(1):1–8.
- Egrioglu, E., Aladag, C. H., Yolcu, U., Uslu, V. R., and Basaran, M. A. (2009). A new approach based on artificial neural networks for high order multivariate fuzzy time series. *Expert Systems with Applications*, 36(7):10589–10594.
- Elliott, G., Rothenberg, T. J., and Stock, J. H. (1992). Efficient tests for an autoregressive unit root. Technical report, National Bureau of Economic Research.
- Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Press.
- Enders, C. K. (2011). Analyzing longitudinal data with missing values. *Rehabilitation Psychology*, 56(4):267.
- Enders, W. (2008). *Applied Econometric Time Series*. John Wiley & Sons.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007.
- Engle, R. F. and Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: Journal of the Econometric Society*, pages 251–276.

References

- Fan, S., Kind, T., Cajka, T., Hazen, S. L., Tang, W. W., Kaddurah-Daouk, R., Irvin, M. R., Arnett, D. K., Barupal, D. K., and Fiehn, O. (2019). Systematic error removal using random forest for normalizing large-scale untargeted lipidomics data. *Analytical Chemistry*, 91(5):3590–3596.
- Fareed, Z., Iqbal, N., Shahzad, F., Shah, S. G. M., Zulfiqar, B., Shahzad, K., Hashmi, S. H., and Shahzad, U. (2020). Co-variance nexus between COVID-19 mortality, humidity, and air quality index in Wuhan, China: New insights from partial and multiple wavelet coherence. *Air Quality, Atmosphere & Health*, 13:673–682.
- Farhangfar, A., Kurgan, L., and Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, 41(12):3692–3705.
- Faust, J. (1996). Near observational equivalence and theoretical size problems with unit root tests. *Econometric Theory*, 12(4):724–731.
- Faustini, A., Rapp, R., and Forastiere, F. (2014). Nitrogen dioxide and mortality: review and meta-analysis of long-term studies. *European Respiratory Journal*, 44(3):744–753.
- Fenger, J. (1999). Urban air quality. *Atmospheric environment*, 33(29):4877–4900.
- Ferreira, T. M., Forti, M. C., De Freitas, C. U., Nascimento, F. P., Junger, W. L., and Gouveia, N. (2016). Effects of particulate matter and its chemical constituents on elderly hospital admissions due to circulatory and respiratory diseases. *International Journal of Environmental Research and Public Health*, 13(10):947.
- Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. Sage.
- Fielding, S., Fayes, P. M., McDonald, A., McPherson, G., Campbell, M. K., Group, R. S., et al. (2008). Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. *Health and Quality of Life Outcomes*, 6(1):57.
- Fitz-Simons, T. (1999). Guideline for reporting of daily air quality: Air Quality Index (AQI). Technical report, Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, NC (United States).

References

- Forbes, D., Hawthorne, G., Elliott, P., McHugh, T., Biddle, D., Creamer, M., and Novaco, R. W. (2004). A concise measure of anger in combat-related posttraumatic stress disorder. *Journal of Traumatic Stress: Official Publication of The International Society for Traumatic Stress Studies*, 17(3):249–256.
- Fransen, J., Creemers, M., and Van Riel, P. (2004). Remission in rheumatoid arthritis: agreement of the disease activity score (DAS28) with the ARA preliminary remission criteria. *Rheumatology*, 43(10):1252–1255.
- Frontera, J. A., Sabadia, S., Lalchan, R., Fang, T., Flusty, B., Millar-Verneti, P., Snyder, T., Berger, S., Yang, D., Granger, A., et al. (2021). A prospective study of neurologic disorders in hospitalized patients with COVID-19 in New York City. *Neurology*, 96(4):e575–e586.
- Fuller, W. A. (2009). *Introduction to Statistical Time Series*. John Wiley & Sons.
- Gabriel, S. E., Crowson, C. S., and O’Fallon, M. (1999). The epidemiology of rheumatoid arthritis in Rochester, Minnesota, 1955–1985. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 42(3):415–420.
- Gauderman, W. J., Gilliland, G. F., Vora, H., Avol, E., Stram, D., McConnell, R., Thomas, D., Lurmann, F., Margolis, H. G., Rappaport, E. B., et al. (2002). Association between air pollution and lung function growth in southern California children: results from a second cohort. *American Journal of Respiratory and Critical Care Medicine*, 166(1):76–84.
- Gautam, S. (2020). COVID-19: air pollution remains low as people stay at home. *Air Quality, Atmosphere & Health*, 13:853–857.
- Ghasemi, A. and Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, 10(2):486.
- Ghazali, S. M., Shaadan, N., and Idrus, Z. (2020). Missing data exploration in air quality

References

- data set using R-package data visualisation tools. *Bulletin of Electrical Engineering and Informatics*, 9(2):755–763.
- Ghorani-Azam, A., Riahi-Zanjani, B., and Balali-Mood, M. (2016). Effects of air pollution on human health and practical measures for prevention in Iran. *Journal of Research in Medical Sciences: the Official Journal of Isfahan University of Medical Sciences*, 21(1):65.
- Girguis, M. S., Li, L., Lurmann, F., Wu, J., Breton, C., Gilliland, F., Stram, D., and Habre, R. (2020). Exposure measurement error in air pollution studies: the impact of shared, multiplicative measurement error on epidemiological health risk estimates. *Air Quality, Atmosphere & Health*, 13:631–643.
- Godha, D., Shi, L., and Mavronicolas, H. (2010). Association between tendency towards depression and severity of rheumatoid arthritis from a national representative sample: the medical expenditure panel survey. *Current Medical Research and Opinion*, 26(7):1685–1690.
- Goldstein, H., Browne, W., and Rasbash, J. (2002). Multilevel modelling of medical data. *Statistics in Medicine*, 21(21):3291–3315.
- Gong, C., Tang, P., and Wang, Y. (2019). Measuring the network connectedness of global stock markets. *Physica A: Statistical Mechanics and its Applications*, 535:122351.
- Gonzalez-Alvaro, I., Ortiz, A., Garcia-Vicuna, R., Balsa, A., Pascual-Salcedo, D., and Laffon, A. (2003). Increased serum levels of interleukin-15 in rheumatoid arthritis with long-term disease. *Clinical and Experimental Rheumatology*, 21(5):639–642.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60:549–576.
- Graham, J. W. (2012). *Missing Data: Analysis and Design*. Springer Science & Business Media.

References

- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438.
- Granger, C. W. and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, 2(2):111–120.
- Granger, C. W. J. et al. (1986). Developments in the study of cointegrated economic variables. In *Oxford Bulletin of Economics and Statistics*. Citeseer.
- Gupta, S., Raghuwanshi, G. S., and Chanda, A. (2020). Effect of weather on COVID-19 spread in the US: A prediction model for India in 2020. *Science of the Total Environment*, 728:138860.
- Guttorp, P. and Sampson, P. D. (1994). 20 methods for estimating heterogeneous spatial covariance functions with environmental applications. *Handbook of Statistics*, 12:661–689.
- Habre, R., Coull, B., Moshier, E., Godbold, J., Grunin, A., Nath, A., Castro, W., Schachter, N., Rohr, A., Kattan, M., et al. (2014). Sources of indoor air pollution in New York City residences of asthmatic children. *Journal of Exposure Science & Environmental Epidemiology*, 24(3):269–278.
- Hadeed, S. J., O'Rourke, M. K., Burgess, J. L., Harris, R. B., and Canales, R. A. (2020). Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of The Total Environment*, 730:139140.
- Haji-Maghsoudi, S., Haghdoost, A.-a., Rastegari, A., and Baneshi, M. R. (2013). Influence of pattern of missing data on performance of imputation methods: an example using national data on drug injection in prisons. *International Journal of Health Policy and Management*, 1(1):69.
- Haldrup, N., Nielsen, F. S., and Nielsen, M. Ø. (2010). A vector autoregressive model for electricity prices subject to long memory and regime switching. *Energy Economics*, 32(5):1044–1058.

References

- Hamilton, J. D. (1994). *Time Series Analysis*, volume 2. Princeton University Press
Princeton, NJ.
- Hamoda, M., Al-Jaralla, R., and Al-Mahamel, S. (2022). Assessment of air pollutants
emissions due to traffic in two residential areas in Kuwait. *International Journal of
Environmental Science and Technology*, 19(2):807–816.
- Hankey, S., Marshall, J. D., and Brauer, M. (2012). Health impacts of the built envi-
ronment: within-urban variability in physical inactivity, air pollution, and ischemic
heart disease mortality. *Environmental Health Perspectives*, 120(2):247–253.
- Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregres-
sion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):190–
195.
- Hart, J. E., Källberg, H., Laden, F., Bellander, T., Costenbader, K. H., Holmqvist, M.,
Klareskog, L., Alfredsson, L., and Karlson, E. W. (2013a). Ambient air pollution
exposures and risk of rheumatoid arthritis: results from the Swedish EIRA case-
control study. *Annals of the Rheumatic Diseases*, 72(6):888–894.
- Hart, J. E., Källberg, H., Laden, F., Costenbader, K. H., Yanosky, J. D., Klareskog, L.,
Alfredsson, L., and Karlson, E. W. (2013b). Ambient air pollution exposures and risk
of rheumatoid arthritis. *Arthritis Care & Research*, 65(7):1190–1196.
- Hart, J. E., Laden, F., Puett, R. C., Costenbader, K. H., and Karlson, E. W. (2009).
Exposure to traffic pollution and increased risk of rheumatoid arthritis. *Environmental
Health Perspectives*, 117(7):1065–1069.
- Hawthorne, G., Hawthorne, G., and Elliott, P. (2005). Imputing cross-sectional miss-
ing data: comparison of common techniques. *Australian & New Zealand Journal of
Psychiatry*, 39(7):583–590.
- Heitjan, D. F. and Rubin, D. B. (1990). Inference from coarse data via multiple imputa-

References

- tion with application to age heaping. *Journal of the American Statistical Association*, 85(410):304–314.
- Henderson, C. R., Kempthorne, O., Searle, S. R., and Von Krosigk, C. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15(2):192–218.
- Higgins, J. P., White, I. R., and Wood, A. M. (2008). Imputation methods for missing outcome data in meta-analysis of clinical trials. *Clinical Trials*, 5(3):225–239.
- Hipel, K. W. and McLeod, A. I. (1994). *Time Series Modelling of Water Resources and Environmental Systems*, volume 45. Elsevier.
- Honaker, J., King, G., Blackwell, M., et al. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47.
- Hondroyannis, G., Lolos, S., and Papapetrou, E. (2002). Energy consumption and economic growth: assessing the evidence from Greece. *Energy Economics*, 24(4):319–336.
- Horton, N. J. and Lipsitz, S. R. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, 55(3):244–254.
- Hossain, M. Z. (2011). The use of box-cox transformation technique in economic and statistical analyses. *Journal of Emerging Trends in Economics and Management Sciences*, 2(1):32–39.
- Huang, K.-H., Yu, T. H.-K., and Hsu, Y. W. (2007). A multivariate heuristic model for fuzzy time series forecasting. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(4):836–846.
- Huisman, M. and Krause, R. W. (2018). Imputation of missing network data. In *Encyclopedia of Social Network Analysis and Mining*, pages 1044–1053. Springer New York.

References

- Hwang, J.-H. and Yoo, S.-H. (2014). Energy consumption, CO_2 emissions, and economic growth: evidence from Indonesia. *Quality & Quantity*, 48(1):63–73.
- Imtiaz, S. and Shah, S. (2008). Treatment of missing values in process data analysis. *The Canadian Journal of Chemical Engineering*, 86(5):838–858.
- Ishak, A. B., Daoud, M. B., and Trabelsi, A. (2017). Ozone concentration forecasting using statistical learning approaches. *Journal of Materials and Environmental Sciences*, 8(12):4532–4543.
- Isufi, E., Loukas, A., Perraudin, N., and Leus, G. (2019). Forecasting time series with VARMA recursions on graphs. *IEEE Transactions on Signal Processing*, 67(18):4870–4885.
- Jadhav, A., Pramod, D., and Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10):913–933.
- Jagannathan, G. and Wright, R. N. (2008). Privacy-preserving imputation of missing data. *Data & Knowledge Engineering*, 65(1):40–56.
- Jain, A., Sukhdeve, T., Gadia, H., Sahu, S. P., and Verma, S. (2021). COVID-19 prediction using time series analysis. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 1599–1606. IEEE.
- Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2-3):231–254.
- Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica: Journal of the Econometric Society*, pages 1551–1580.
- Johansen, S. (1992). Cointegration in partial systems and the efficiency of single-equation analysis. *Journal of Econometrics*, 52(3):389–402.

References

- Johansen, S. (1995). *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*. OUP Oxford.
- Johansen, S. (2000). Modelling of cointegration in the vector autoregressive model. *Economic Modelling*, 17(3):359–373.
- Johansen, S. and Juselius, K. (1990). Maximum likelihood estimation and inference on cointegration—with applications to the demand for money. *Oxford Bulletin of Economics and Statistics*, 52(2):169–210.
- Johnson, D. L., Ambrose, S. H., Bassett, T. J., Bowen, M. L., Crummey, D. E., Isaacson, J. S., Johnson, D. N., Lamb, P., Saul, M., and Winter-Nelson, A. E. (1997). Meanings of environmental terms. *Journal of Environmental Quality*, 26(3):581–589.
- Johnson, M., Isakov, V., Touma, J., Mukerjee, S., and Özkaynak, H. (2010). Evaluation of land-use regression models used to predict air quality concentrations in an urban area. *Atmospheric Environment*, 44(30):3660–3668.
- Jones, S. S., Thomas, A., Evans, R. S., Welch, S. J., Haug, P. J., and Snow, G. L. (2008). Forecasting daily patient volumes in the emergency department. *Academic Emergency Medicine*, 15(2):159–170.
- Jung, C.-R., Hsieh, H.-Y., and Hwang, B.-F. (2017). Air pollution as a potential determinant of rheumatoid arthritis: a population-based cohort study in Taiwan. *Epidemiology*, 28:S54–S59.
- Junger, W. and de Leon, A. P. (2009). Missing data imputation in time series of air pollution. *Epidemiology*, 20(6):S87.
- Junger, W. and De Leon, A. P. (2015). Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, 102:96–104.
- Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., and Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, 38(18):2895–2907.

References

- Juselius, K. (2006). *The Cointegrated VAR Model: Methodology and Applications*. Oxford University Press.
- Kabir, G., Tesfamariam, S., Hemsing, J., and Sadiq, R. (2019). Handling incomplete and missing data in water network database using imputation methods. *Sustainable and Resilient Infrastructure*, 5(6):365–377.
- Kanchan, A. K. G., Goyal, P., Benitez-Garcia, S. E., Kanda, I., Okazaki, Y., Wakamatsu, S., Basaldud, R., Horikoshi, N., Ortinez, J. A., Ramos-Benitez, V. R., et al. (2015). A review on air quality indexing system. *Asian Journal of Atmospheric Environment*, 9(2):101–113.
- Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5):402.
- Karimi, A., Shirmardi, M., Hadei, M., Birgani, Y. T., Neisi, A., Takdastan, A., and Goudarzi, G. (2019). Concentrations and health effects of short-and long-term exposure to $PM_{2.5}$, NO_2 , and O_3 in ambient air of Ahvaz City, Iran (2014–2017). *Ecotoxicology and Environmental Safety*, 180:542–548.
- Kaufmann, R. K., Kauppi, H., and Stock, J. H. (2006). The relationship between radiative forcing and temperature: what do statistical analyses of the instrumental temperature record measure? *Climatic Change*, 77(3):279–289.
- Kayes, I., Shahriar, S., Hasan, K., Akhter, M., Kabir, M., Salam, M., et al. (2019). The relationships between meteorological parameters and air pollutants in an urban environment. *Global J. Environ. Sci. Management*, 5(3):265–278.
- Khaniabadi, Y. O., Sicard, P., Khaniabadi, A. O., Mohammadinejad, S., Keishams, F., Takdastan, A., Najafi, A., De Marco, A., and Daryanoosh, M. (2018). Air quality modeling for health risk assessment of ambient pm10, pm2. 5 and so2 in iran. *Human and Ecological Risk Assessment: An International Journal*.

References

- Khin, A. A., Thambiah, S., and Pushpakumara, D. (2015). Natural rubber prices forecasting using simultaneous supply-demand and price system equation and VECM model: Between theory and reality. In *Proceedings of 2nd International Conference on Agriculture and Forestry, ICOAF-2015, Colombo, Sri Lanka*, pages 184–195.
- Khorsandi, B., Farzad, K., Tahri, H., and Maknoon, R. (2021). Association between short-term exposure to air pollution and COVID-19 hospital admission/mortality during warm seasons. *Environmental Monitoring and Assessment*, 193(7):1–6.
- King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95(1):49–69.
- Kinney, P. L., Aggarwal, M., Northridge, M. E., Janssen, N. A., and Shepard, P. (2000). Airborne concentrations of $PM_{2.5}$ and diesel exhaust particles on Harlem sidewalks: a community-based pilot study. *Environmental Health Perspectives*, 108(3):213–218.
- Kobayashi, S., Okamoto, H., Iwamoto, T., Toyama, Y., Tomatsu, T., Yamanaka, H., and Momohara, S. (2008). A role for the aryl hydrocarbon receptor and the dioxin TCDD in rheumatoid arthritis. *Rheumatology*, 47(9):1317–1322.
- Kočenda, E. and Černý, A. (2015). *Elements of Time Series Econometrics: An Applied Approach*. Charles University in Prague, Karolinum Press.
- Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J., and Hanhineva, K. (2019). Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. *BMC Bioinformatics*, 20(1):1–11.
- Konarasinghe, K. (2020). Modeling COVID-19 epidemic of USA, UK and Russia. *Journal of New Frontiers in Healthcare and Biological Sciences*, 1(1):1–14.
- Korkas, K. K. and Pryzlewicz, P. (2017). Multiple change-point detection for non-stationary time series using wild binary segmentation. *Statistica Sinica*, 27(1):287–311.

References

- Kotsiantis, S., Kostoulas, A., Lykoudis, S., Argiriou, A., and Menagias, K. (2006). Filling missing temperature values in weather data banks. In *2006 2nd IET International Conference on Intelligent Environments-IE 06*, volume 1, pages 327–334. IET.
- Kowalska, M., Ośródk, L., Klejnowski, K., Zejda, J. E., Krajny, E., and Wojtylak, M. (2009). Air quality index and its significance in environmental health risk communication. *Archives of Environmental Protection*, 35(1):13–21.
- Kowarik, A. and Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software*, 74(7):1–16.
- Kumar, A. and Goyal, P. (2011). Forecasting of daily air quality index in Delhi. *Science of the Total Environment*, 409(24):5517–5523.
- Kumar, U. and Jain, V. (2010). ARIMA forecasting of ambient air pollutants (o₃, no, no₂ and co). *Stochastic Environmental Research and Risk Assessment*, 24(5):751–760.
- Kumari, B. et al. (2018). Effects of air pollution on health and environment. *International Journal for Research in Applied Sciences and Biotechnology*, 5(6):26–33.
- Künzli, N., Kaiser, R., Medina, S., Studnicka, M., Chanel, O., Filliger, P., Herry, M., Horak Jr, F., Puybonnieux-Textier, V., Quénel, P., et al. (2000). Public-health impact of outdoor and traffic-related air pollution: a european assessment. *The Lancet*, 356(9232):795–801.
- Lakshminarayan, K., Harp, S. A., and Samad, T. (1999). Imputation of missing data in industrial databases. *Applied Intelligence*, 11(3):259–275.
- Lanne, M. and Nyberg, H. (2016). Generalized forecast error variance decomposition for linear and nonlinear multivariate models. *Oxford Bulletin of Economics and Statistics*, 78(4):595–603.
- Latief, R., Kong, Y., Javeed, S. A., and Sattar, U. (2021). Carbon emissions in the SAARC countries with causal effects of FDI, economic growth and other economic

References

- factors: Evidence from dynamic simultaneous equation models. *International Journal of Environmental Research and Public Health*, 18(9):4605.
- Lau, L.-S., Yip, K.-J., Lee, C.-Y., Chong, Y.-L., and Lee, E.-H. (2018). Investigating the determinants of renewable energy consumption in Malaysia: an ARDL approach. *International Journal of Business and Society*, 19(3):886–903.
- Lauc, G., Markotić, A., Gornik, I., and Primorac, D. (2020). Fighting COVID-19 with water. *Journal of Global Health*, 10(1):010344.
- Lee, K. J. and Carlin, J. B. (2012). Recovery of information from multiple imputation: a simulation study. *Emerging Themes in Epidemiology*, 9(1):3.
- Li, H., Li, X., Zheng, H., Liu, L., Wu, Y., Zhou, Y., Meng, X., Hong, J., Cao, L., Lu, Y., et al. (2021). Ultrafine particulate air pollution and pediatric emergency-department visits for main respiratory diseases in Shanghai, China. *Science of The Total Environment*, 775:145777.
- Li, P., Stuart, E. A., and Allison, D. B. (2015). Multiple imputation: a flexible tool for handling missing data. *JAMA*, 314(18):1966–1967.
- Li, W., Dorans, K. S., Wilker, E. H., Rice, M. B., Kloog, I., Schwartz, J. D., Koutrakis, P., Coull, B. A., Gold, D. R., Meigs, J. B., et al. (2018). Ambient air pollution, adipokines, and glucose homeostasis: The Framingham heart study. *Environment International*, 111:14–22.
- Liang, C.-S., Duan, F.-K., He, K.-B., and Ma, Y.-L. (2016). Review on recent progress in observations, source identifications and countermeasures of PM_{10} . *Environment International*, 86:150–170.
- Liao, S. G., Lin, Y., Kang, D. D., Chandra, D., Bon, J., Kaminski, N., Sciruba, F. C., and Tseng, G. C. (2014). Missing value imputation in high-dimensional phenomic data: imputable or not, and how? *BMC Bioinformatics*, 15(1):346.

References

- Libasin, Z., Ul-Saufie, A. Z., Ahmat, H., and Shaziayani, W. N. (2020). Single and multiple imputation method to replace missing values in air pollution datasets: A review. In *IOP Conference Series: Earth and Environmental Science*, volume 616, page 012002. IOP Publishing.
- Lim, S. S., Vos, T., Flaxman, A. D., Danaei, G., Shibuya, K., Adair-Rohani, H., Al-Mazroa, M. A., Amann, M., Anderson, H. R., Andrews, K. G., et al. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010. *The Lancet*, 380(9859):2224–2260.
- Lin, S., Liu, X., Le, L. H., and Hwang, S.-A. (2008). Chronic exposure to ambient ozone and asthma hospital admissions among children. *Environmental Health Perspectives*, 116(12):1725–1730.
- Little, R. J. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237.
- Little, R. J. and Rubin, D. B. (2002). *Bayes and multiple imputation*. John Wiley Sons, Inc.
- Little, R. J. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*, volume 793. John Wiley Sons, Inc., USA.
- Liu, H. and Rodríguez, G. (2005). Human activities and global warming: a cointegration analysis. *Environmental Modelling & Software*, 20(6):761–773.
- Liu, J., Zhou, J., Yao, J., Zhang, X., Li, L., Xu, X., He, X., Wang, B., Fu, S., Niu, T., et al. (2020). Impact of meteorological factors on the COVID-19 transmission: A multi-city study in China. *Science of the Total Environment*, 726:138513.
- Ljung, G. M. and Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.

References

- Loganathan, N. and Subramaniam, T. (2010). Dynamic cointegration link between energy consumption and economic performance: empirical evidence from Malaysia. *International Journal of Trade, Economics and Finance*, 1(3):261–267.
- Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media.
- Lütkepohl, H. (2010). Impulse response function. In *Macroeconometrics and Time Series Analysis*, pages 145–150. Springer.
- Lütkepohl, H. (2013). *Introduction to Multiple Time Series Analysis*. Springer Science & Business Media.
- Ma, Y., Zhao, Y., Liu, J., He, X., Wang, B., Fu, S., Yan, J., Niu, J., Zhou, J., and Luo, B. (2020). Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. *Science of the Total Environment*, 724:138226.
- Machol, B. and Rizk, S. (2013). Economic value of us fossil fuel electricity health impacts. *Environment international*, 52:75–80.
- MacKinnon, J. G. (1996). Numerical distribution functions for unit root and cointegration tests. *Journal of Applied Econometrics*, 11(6):601–618.
- MacNally, R. (2000). Regression and model-building in conservation biology, biogeography and ecology: the distinction between—and reconciliation of—‘predictive’and ‘explanatory’models. *Biodiversity & Conservation*, 9(5):655–671.
- Malarvizhi, R. and Thanamani, A. S. (2012). K-nearest neighbor in missing data imputation. *International Journal of Engineering Research and Development*, 5(1):5–7.
- Malig, B. J., Pearson, D. L., Chang, Y. B., Broadwin, R., Basu, R., Green, R. S., and Ostro, B. (2016). A time-stratified case-crossover study of ambient ozone exposure and emergency department visits for specific respiratory diagnoses in California (2005–2008). *Environmental Health Perspectives*, 124(6):745–753.

References

- Manisalidis, I., Stavropoulou, E., Stavropoulos, A., and Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: a review. *Frontiers in Public Health*, 8:14.
- Marshall, A., Altman, D. G., and Holder, R. L. (2010). Comparison of imputation methods for handling missing covariate data when fitting a cox proportional hazards model: a resampling study. *BMC Medical Research Methodology*, 10(1):112.
- Martelletti, L. and Martelletti, P. (2020). Air pollution and the novel COVID-19 disease: a putative disease risk factor. *SN Comprehensive Clinical Medicine*, 2(4):383–387.
- Martínez, G., Feist, E., Martiatu, M., Garay, H., and Torres, B. (2020). Autoantibodies against a novel citrullinated fibrinogen peptide related to smoking status, disease activity and therapeutic response to methotrexate in cuban patients with early rheumatoid arthritis. *Rheumatology International*, 40:1873–1881.
- Martins, F. M., da Silva, J. A. P., Santos, M. J., Vieira-Sousa, E., Duarte, C., Santos, H., Costa, J. A., Pimentel-Santos, F. M., Cunha, I., Cunha Miranda, L., et al. (2014). Das28, cdai and sdai cut-offs do not translate the same information: results from the Rheumatic Diseases Portuguese Register Reuma. pt. *Rheumatology*, 54(2):286–291.
- Masri, S., Garshick, E., Hart, J., Bouhamra, W., and Koutrakis, P. (2017). Use of visual range measurements to predict fine particulate matter exposures in southwest Asia and Afghanistan. *Journal of the Air & Waste Management Association*, 67(1):75–85.
- McKnight, P. E., McKnight, K. M., Sidani, S., and Figueredo, A. J. (2007). *Missing Data: A Gentle Introduction*. Guilford Press.
- McNeish, D. (2017). Missing data methods for arbitrary missingness with small samples. *Journal of Applied Statistics*, 44(1):24–39.
- Menebo, M. M. (2020). Temperature and precipitation associate with COVID-19 new daily cases: A correlation study between weather and COVID-19 pandemic in Oslo, Norway. *Science of the Total Environment*, 737:139659.

References

- Mintz, D. (2006). Guidelines for the reporting of daily air quality—air quality index (AQI). *Washington: United States Environmental Protection Agency.*
- Mintz, D. (2009). Technical assistance document for the reporting of daily air quality—the air quality index (aqi). *Tech. Research Triangle Park, US Environmental Protection Agency.*
- Mirsaeidi, M., Motahari, H., Taghizadeh Khamesi, M., Sharifi, A., Campos, M., and Schraufnagel, D. E. (2016). Climate change and respiratory infections. *Annals of the American Thoracic Society*, 13(8):1223–1230.
- Mishra, S. and Khare, D. (2014). On comparative performance of multiple imputation methods for moderate to large proportions of missing data in clinical trials: a simulation study. *Journal of Medical Statistics and Informatics*, 2(1):9.
- Misztal, M. (2013). Some remarks on the data imputation using “missforest” method.
- Molenberghs, G. and Kenward, M. (2007). *Missing Data in Clinical Studies*. John Wiley & Sons.
- Mölder, A., Simpson, A., Berdel, D., Brunekreef, B., Custovic, A., Cyrus, J., de Jongste, J., De Vocht, F., Fuertes, E., Gehring, U., et al. (2015). A multicentre study of air pollution exposure and childhood asthma prevalence: the ESCAPE project. *European Respiratory Journal*, 45(3):610–624.
- Moons, K. G., Donders, R. A., Stijnen, T., and Harrell Jr, F. E. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, 59(10):1092–1101.
- Mosley, S. (2014). Environmental history of air pollution and protection. In *The basic environmental history*, pages 143–169. Springer.
- Muñoz, J. G. B., Giraldo, R. B., Santos, A. M., Bello-Gualteros, J. M., Rueda, J. C., Saldarriaga, E.-L., Angarita, J.-I., Arias-Correal, S., Vasquez, A. Y., and Londono,

References

- J. (2017). Correlation between rapid-3, DAS28, CDAI and SDAI as a measure of disease activity in a cohort of colombian patients with rheumatoid arthritis. *Clinical Rheumatology*, 36(5):1143–1148.
- Murena, F. (2004). Measuring air quality over large urban areas: development and application of an air pollution index at the urban area of Naples. *Atmospheric Environment*, 38(36):6195–6202.
- Murugesan, B., Karuppanan, S., Mengistie, A. T., Ranganathan, M., and Gopalakrishnan, G. (2020). Distribution and trend analysis of COVID-19 in India: geospatial approach. *J Geogr Stud*, 4(1):1–9.
- Mustafa, H. I. and Fareed, N. Y. (2020). COVID-19 cases in Iraq; forecasting incidents using Box-Jenkins ARIMA model. In *2020 2nd Al-Noor International Conference for Science and Technology (NICST)*, pages 22–26. IEEE.
- Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, 135(3):370–384.
- Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, 17(4):372–411.
- Nguyen, H. M., Turk, P., and McWilliams, A. (2021). A multivariate forecasting model for the COVID-19 hospital census based on local infection incidence. *medRxiv*.
- Nguyen, N. P. and Marshall, J. D. (2018). Impact, efficiency, inequality, and injustice of urban air pollution: Variability by emission location. *Environmental Research Letters*, 13(2):024002.
- Niu, Z., Liu, F., Yu, H., Wu, S., and Xiang, H. (2021). Association between exposure to ambient air pollution and hospital admission, incidence, and mortality of stroke: an updated systematic review and meta-analysis of more than 23 million participants. *Environmental Health and Preventive Medicine*, 26(1):1–14.

References

- Norazian, M. N., Shukri, Y. A., Azam, R. N., and Al Bakri, A. M. M. (2008). Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia*, 34(3):341–345.
- Oba, S., Sato, M.-a., Takemasa, I., Monden, M., Matsubara, K.-i., and Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096.
- Ogen, Y. (2020). Assessing nitrogen dioxide (NO_2) levels as a contributing factor to coronavirus (COVID-19) fatality. *Science of the Total Environment*, 726:138605.
- Organization, W. H. (2006). *Air Quality Guidelines: Global Update 2005: Particulate Matter, Ozone, Nitrogen Dioxide, and Sulfur Dioxide*. World Health Organization.
- Organization, W. H. (2016). *Ambient Air Pollution: a Global Assessment of Exposure and Burden of Disease*. World Health Organization.
- Organization, W. H. et al. (1997). Intersectoral action for health: Addressing health and environment concerns in sustainable development. Technical report, World Health Organization.
- Organization, W. H. et al. (1999). Air quality management: Air quality guidelines. *World Health Organization, Geneva (available as of 5 January 2000 at: www.who.int/peh/air/airqualitygd.htm)*.
- Organization, W. H. et al. (2005). WHO air quality guidelines global update 2005: Report on a working group meeting, Bonn, Germany, 18–20 october 2005. Technical report, World Health Organization. Regional Office for Europe.
- Organization, W. H. et al. (2021). *WHO global air quality guidelines: particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization.

References

- Pahlavani, M., Wilson, E., and Worthington, A. C. (2005). Trade-gdp nexus in Iran: An application of the autoregressive distributed lag (ARDL) model. *American Journal of Applied Sciences*, 2(7):1158–1165.
- Paik, M. C. and Sacco, R. L. (2000). Matched case-control data analyses with missing covariates. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(1):145–156.
- Pani, S. K., Lin, N.-H., and RavindraBabu, S. (2020). Association of COVID-19 pandemic with meteorological parameters over Singapore. *Science of the Total Environment*, 740:140112.
- Pata, U. K. (2020). How is COVID-19 affecting environmental pollution in us cities? evidence from asymmetric fourier causality test. *Air Quality, Atmosphere & Health*, 13(10):1149–1155.
- Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., and Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, 9:157.
- Peng, L., Zhao, X., Tao, Y., Mi, S., Huang, J., and Zhang, Q. (2020). The effects of air pollution and meteorological factors on measles cases in Lanzhou, China. *Environmental Science and Pollution Research*, 27(12):13524–13533.
- Pereira, P. L. V. et al. (2004). How persistent is volatility? an answer with stochastic volatility models with markov regime switching state equations. Technical report, Citeseer.
- Perron, P. (1990). Testing for a unit root in a time series with a changing mean. *Journal of Business & Economic Statistics*, 8(2):153–162.
- Peters, A., Perz, S., Döring, A., Stieber, J., Koenig, W., and Wichmann, H. E. (1999). Increases in heart rate during an air pollution episode. *American journal of epidemiology*, 150(10):1094–1098.

References

- Pfaff, B. et al. (2008). VAR, SVAR and SVEC models: Implementation within R package vars. *Journal of Statistical Software*, 27(4):1–32.
- Phillips, P. C. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75(2):335–346.
- Pirouz, B., Golmohammadi, A., Masouleh, H. S., Delazzari, C., Violini, G., and Pirouz, B. (2020). Relationship between average daily temperature and average cumulative daily rate of confirmed cases of COVID-19. *medRxiv*.
- Plaia, A. and Bondi, A. (2006). Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment*, 40(38):7316–7330.
- Plaia, A. and Ruggieri, M. (2011). Air quality indices: a review. *Reviews in Environmental Science and Bio/Technology*, 10(2):165–179.
- Polit, D. F. and Beck, C. T. (2008). *Nursing Research: Generating and Assessing Evidence for Nursing Practice*. Lippincott Williams & Wilkins.
- Pope, C. A., Thun, M. J., Namboodiri, M. M., Dockery, D. W., Evans, J. S., Speizer, F. E., Heath, C. W., et al. (1995). Particulate air pollution as a predictor of mortality in a prospective study of US adults. *American Journal of Respiratory and Critical Care Medicine*, 151(3):669–674.
- Powe, N. A. and Willis, K. G. (2004). Mortality and morbidity benefits of air pollution (so₂ and pm₁₀) absorption attributable to woodland in Britain. *Journal of environmental management*, 70(2):119–128.
- Powell, H. L. (2012). *Estimating air pollution and its relationship with human health*. PhD thesis, University of Glasgow.
- Prevoo, M., Van’T Hof, M., Kuper, H., Van Leeuwen, M., Van De Putte, L., and Van Riel, P. (1995). Modified disease activity scores that include twenty-eight-joint counts development and validation in a prospective longitudinal study of patients

References

- with rheumatoid arthritis. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 38(1):44–48.
- Prospero, J., Barrett, K., Church, T., Dentener, F., Duce, R., Galloway, J., Levy, H., Moody, J., and Quinn, P. (1996). Atmospheric deposition of nutrients to the North Atlantic Basin. In *Nitrogen Cycling in the North Atlantic Ocean and its Watersheds*, pages 27–73. Springer.
- Pruss-Ustun, A., Corvalán, C. F., Organization, W. H., et al. (2006). *Preventing Disease Through Healthy Environments: Towards an Estimate of the Environmental Burden of Disease*. World Health Organization.
- Raicharoen, T., Lursinsap, C., and Sanguanbhokai, P. (2003). Application of critical support vector machine to time series prediction. In *Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS'03.*, volume 5, pages V–V. IEEE.
- Rao, S., Klimont, Z., Smith, S. J., Van Dingenen, R., Dentener, F., Bouwman, L., Riahi, K., Amann, M., Bodirsky, B. L., van Vuuren, D. P., et al. (2017). Future air pollution in the shared socio-economic pathways. *Global Environmental Change*, 42:346–358.
- Richards, A. J. (1995). Comovements in national stock market returns: Evidence of predictability, but not cointegration. *Journal of Monetary Economics*, 36(3):631–654.
- Robinson, D. L. (2005). Air pollution in Australia: review of costs, sources and potential solutions. *Health Promotion Journal of Australia*, 16(3):213–220.
- Rodríguez-Camargo, L. A., Sierra-Parada, R. J., and Blanco-Becerra, L. C. (2020). Spatial analysis of $PM_{2.5}$ concentrations in Bogotá according to the World Health Organization air quality guidelines for cardiopulmonary diseases, 2014-2015. *Biomédica*, 40(1):137–152.
- Roy, S., Bhunia, G. S., and Shit, P. K. (2021). Spatial prediction of COVID-19 epi-

References

- demic using ARIMA techniques in India. *Modeling Earth Systems and Environment*, 7(2):1385–1391.
- Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, 4(3):227–241.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.
- Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley Sons, Inc., USA.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.
- Sacramento, D. S., Martins, L. C., Arbex, M. A., and Pamplona, Y. d. A. (2020). Atmospheric pollution and hospitalization for cardiovascular and respiratory diseases in the City of Manaus from 2008 to 2012. *The Scientific World Journal*, 2020:8458359–8.
- Sahai, A. K., Rath, N., Sood, V., and Singh, M. P. (2020). ARIMA modelling & forecasting of COVID-19 in top five affected countries. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(5):1419–1427.
- Said, S. E. and Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607.
- Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(2):169–178.
- Salaffi, F., Cimmino, M., Leardini, G., Gasparini, S., Grassi, W., et al. (2009). Disease activity assessment of rheumatoid arthritis in daily practice: validity, internal consistency, reliability and congruency of the disease activity score including 28 joints (DAS28) compared with the clinical disease activity index (CDAI). *Clinical and Experimental Rheumatology*, 27(4):552–9.

References

- Sartori, N., Salvan, A., and Thomaseth, K. (2005). Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose. *Computational Statistics & Data Analysis*, 49(3):937–953.
- Saxena, P. and Srivastava, A. (2020). *Air Pollution and Environmental Health*. Springer.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC.
- Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2):147.
- Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst’s perspective. *Multivariate Behavioral Research*, 33(4):545–571.
- Schenker, N. and Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, 22(4):425–446.
- Schwartz, J. and Marcus, A. (1990). Mortality and air pollution in London: A time series analysis. *American Journal of Epidemiology*, 131(1):185–194.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.
- Scott, H. M., Soskolne, C. L., Martin, S. W., Ellehoj, E. A., Coppock, R. W., Guidotti, T. L., and Lissemore, K. D. (2003). Comparison of two atmospheric-dispersion models to assess farm-site exposure to sour-gas processing-plant emissions. *Preventive Veterinary Medicine*, 57(1-2):15–34.
- Seal, H. L. (1967). Studies in the history of probability and statistics. XV the historical development of the Gauss linear model. *Biometrika*, 54(1-2):1–24.
- Seinfeld, J. H. (1975). *Air Pollution: Physical and Chemical Fundamentals*. McGraw-Hill.

References

- Seinfeld, J. H. (1989). Urban air pollution: state of the science. *Science*, 243(4892):745–752.
- Sephton, P. S. (1995). Response surface estimates of the KPSS stationarity test. *Economics Letters*, 47(3-4):255–261.
- Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American Journal of Epidemiology*, 179(6):764–774.
- Shahn, Z., Ryan, P., and Madigan, D. (2015). Predicting health outcomes from high-dimensional longitudinal health histories using relational random forests. *Statistical Analysis and Data Mining: the ASA Data Science Journal*, 8(2):128–136.
- Sharma, S., Zhang, M., Gao, J., Zhang, H., Kota, S. H., et al. (2020). Effect of restricted emissions during COVID-19 on air quality in India. *Science of the Total Environment*, 728:138878.
- Shehzad, K., Sarfraz, M., and Shah, S. G. M. (2020). The impact of COVID-19 as a necessary evil on air pollution in india during the lockdown. *Environmental Pollution*, 266:115080.
- Shen, L. and Mickley, L. J. (2017). Effects of el niño on summertime ozone air quality in the eastern United States. *Geophysical Research Letters*, 44(24):12–543.
- Shen, L., Zhang, H., Zhou, X., and Liu, R. (2015). Association between polymorphisms of interleukin 12 and rheumatoid arthritis associated biomarkers in a Chinese population. *Cytokine*, 76(2):363–367.
- Shi, P., Dong, Y., Yan, H., Zhao, C., Li, X., Liu, W., He, M., Tang, S., and Xi, S. (2020). Impact of temperature on the dynamics of the COVID-19 outbreak in China. *Science of the Total Environment*, 728:138890.

References

- Shin, J., Lee, J., Lee, J., and Ha, E.-H. (2019). Association between exposure to ambient air pollution and rheumatoid arthritis in adults. *International Journal of Environmental Research and Public Health*, 16(7):1227.
- Shrive, F. M., Stuart, H., Quan, H., and Ghali, W. A. (2006). Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. *BMC medical research methodology*, 6(1):1–10.
- Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, 48(1):1–48.
- Smith, K. R., Jerrett, M., Anderson, H. R., Burnett, R. T., Stone, V., Derwent, R., Atkinson, R. W., Cohen, A., Shonkoff, S. B., Krewski, D., et al. (2009). Public health benefits of strategies to reduce greenhouse-gas emissions: health implications of short-lived greenhouse pollutants. *The lancet*, 374(9707):2091–2103.
- Smith, T., Blumenthal, D., Anderson, J., and Vanderpol, A. (1978). Transport of so₂ in power plant plumes: Day and night. In *Sulfur in the Atmosphere*, pages 605–611. Elsevier.
- Smolen, J., Breedveld, F., Schiff, M., Kalden, J., Emery, P., Eberl, G., Van Riel, P., and Tugwell, P. (2003). A simplified disease activity index for rheumatoid arthritis for use in clinical practice. *Rheumatology*, 42(2):244–257.
- Sobral, M. F. F., Duarte, G. B., da Penha Sobral, A. I. G., Marinho, M. L. M., and de Souza Melo, A. (2020). Association between climate variables and global transmission of SARS-CoV-2. *Science of The Total Environment*, 729:138997.
- Solus, J. F., Chung, C. P., Oeser, A., Li, C., Rho, Y. H., Bradley, K. M., Kawai, V. K., Smith, J. R., and Stein, C. M. (2015). Genetics of serum concentration of IL-6 and TNF α in systemic lupus erythematosus and rheumatoid arthritis: a candidate gene analysis. *Clinical Rheumatology*, 34(8):1375–1382.

References

- Sparks, J. J. and Yurova, Y. V. (2006). Comparative performance of ARIMA and ARCH/GARCH models on time series of daily equity prices for large companies. *2006 SWDSI Proceedings*, pages 563–573.
- Stacklies, W., Redestig, H., Wright, K., Rcpp, S., and Non, I. (2016). Package ‘pcamethods’.
- Stavseth, M. R., Clausen, T., and Røislien, J. (2019). How handling missing data may impact conclusions: a comparison of six different imputation methods for categorical questionnaire data. *SAGE Open Medicine*, 7:2050312118822912.
- Steinfeld, J. I. (1998). Atmospheric chemistry and physics: from air pollution to climate change. *Environment: Science and Policy for Sustainable Development*, 40(7):26–26.
- Stekhoven, D. J. and Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.
- Stern, D. I. and Kaufmann, R. K. (1999). Econometric analysis of global climate change. *Environmental Modelling & Software*, 14(6):597–605.
- Stern, D. I. and Kaufmann, R. K. (2000). Detecting a global warming signal in hemispheric temperature series: A structural time series analysis. *Climatic Change*, 47(4):411–438.
- Stieb, D. M., Doiron, M. S., Blagden, P., and Burnett, R. T. (2005). Estimating the public health burden attributable to air pollution: an illustration using the development of an alternative air quality index. *Journal of Toxicology and Environmental Health, Part A*, 68(13-14):1275–1288.
- Stock, J. H. and Watson, M. W. (2001). Vector autoregressions. *Journal of Economic Perspectives*, 15(4):101–115.
- Sulasikin, A., Nugraha, Y., Kanggrawan, J., and Suherman, A. L. (2020). Forecasting for a data-driven policy using time series methods in handling COVID-19 pandemic in Jakarta. In *The 6th IEEE International Smart Cities Conference (ISC2 2020)*.

References

- Sun, G., Hazlewood, G., Bernatsky, S., Kaplan, G. G., Eksteen, B., and Barnabe, C. (2016). Association between air pollution and the development of rheumatic disease: a systematic review. *International Journal of Rheumatology*, 2016:1–11.
- Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958.
- Taghizadeh-Hesary, F. and Taghizadeh-Hesary, F. (2020). The impacts of air pollution on health and economy in Southeast Asia. *Energies*, 13(7):1812.
- Tang, F. and Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377.
- Tang, J., Yuan, X., Ramos, V., and Sriboonchitta, S. (2019). Does air pollution decrease inbound tourist arrivals? the case of Beijing. *Asia Pacific Journal of Tourism Research*, 24(6):597–605.
- Team, R. C. (2014). R: A language and environment for statistical computing (version 2.15. 2)[computer software]. Vienna, Austria: R foundation for statistical computing.
- Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics*, 3(2):245–265.
- Tian, Y., Liu, H., Zhao, Z., Xiang, X., Li, M., Juan, J., Song, J., Cao, Y., Wang, X., Chen, L., et al. (2018). Association between ambient air pollution and daily hospital admissions for ischemic stroke: a nationwide time series analysis. *PLoS Medicine*, 15(10):e1002668.
- Tobón, G. J., Youinou, P., and Saraux, A. (2010). The environment, geo-epidemiology, and autoimmune disease: Rheumatoid arthritis. *Autoimmunity Reviews*, 9(5):A288–A292.

References

- Tosepu, R., Gunawan, J., Effendy, D. S., Lestari, H., Bahar, H., Asfian, P., et al. (2020). Correlation between weather and COVID-19 pandemic in Jakarta, Indonesia. *Science of The Total Environment*, 725:138436.
- Tsiampalis, T. and Panagiotakos, D. B. (2020). Missing-data analysis: socio-demographic, clinical and lifestyle determinants of low response rate on self-reported psychological and nutrition related multi-item instruments in the context of the AT-TICA epidemiological study. *BMC Medical Research Methodology*, 20(1):1–13.
- Tsiouri, V., Kakosimos, K. E., and Kumar, P. (2015). Concentrations, sources and exposure risks associated with particulate matter in the Middle East area—a review. *Air Quality, Atmosphere & Health*, 8(1):67–80.
- Tsoeleng, L. T. and Shikwambana, L. (2020). Impacts of population growth and land use on air quality. a case study of tshwane, rustenburg and emalahleni, south africa. *South African Geographical Journal= Suid-Afrikaanse Geografiese Tydskrif*, 102(2):209–222.
- Tyagi, R., Bramhankar, M., Pandey, M., and Kishore, M. (2020). COVID 19: Real-time forecasts of confirmed cases, active cases, and health infrastructure requirements for India and its majorly affected states using the ARIMA model. *medRxiv*.
- Valdiviezo, H. C. and Van Aelst, S. (2015). Tree-based prediction on incomplete data using imputation or surrogate decisions. *Information Sciences*, 311:163–181.
- Vallejo, L. A. M., Pardo, M. A. H., Piracón, J. A. B., Cerón, L. C. B., and Achury, N. J. M. (2021). Exposure levels to $PM_{2.5}$ and black carbon for people with disabilities in rural homes of Colombia. *Environmental Monitoring and Assessment*, 193(1):1–19.
- Van Buuren (2018). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press LLC.
- Van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., Jolani, S., et al. (2015). *Package ‘mice’*. Vienna, Austria, 2.22 edition.

References

- Van den Elshout, S., Léger, K., and Heich, H. (2014). Caqi common air quality index—update with PM_{10} and sensitivity analysis. *Science of the Total Environment*, 488:461–468.
- Van den Elshout, S., Léger, K., and Nussio, F. (2008). Comparing urban air quality in Europe in real time: A review of existing air quality indices and the proposal of a common alternative. *Environment International*, 34(5):720–726.
- Van der Heijden, G. J., Donders, A. R. T., Stijnen, T., and Moons, K. G. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *Journal of Clinical Epidemiology*, 59(10):1102–1109.
- Van Gestel, A. M., Haagsma, C. J., and van Riel, P. L. (1998). Validation of rheumatoid arthritis improvement criteria that include simplified joint counts. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 41(10):1845–1850.
- Van Ginkel, J. R., Van der Ark, L. A., Sijtsma, K., and Vermunt, J. K. (2007). Two-way imputation: A Bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation. *Computational Statistics & Data Analysis*, 51(8):4013–4027.
- Van Riel, P. (2014). The development of the disease activity score (DAS) and the disease activity score using 28 joint counts (DAS28). *Clin Exp Rheumatol*, 32(5 Suppl 85):S65–74.
- Vardoulakis, S., Fisher, B. E., Pericleous, K., and Gonzalez-Flesca, N. (2003). Modelling air quality in street canyons: a review. *Atmospheric Environment*, 37(2):155–182.
- Veenstra, A. W. and Haralambides, H. E. (2001). Multivariate autoregressive models for forecasting seaborne trade flows. *Transportation Research Part E: Logistics and Transportation Review*, 37(4):311–319.

References

- Violato, M., Petrou, S., and Gray, R. (2009). The relationship between household income and childhood respiratory health in the United Kingdom. *Social Science & Medicine*, 69(6):955–963.
- Wang, W. (2006). *Stochasticity, Nonlinearity and Forecasting of Stream Flow Processes*. Ios Press.
- Wang, W. and Guo, Y. (2009). Air pollution PM2.5 data analysis in Los Angeles Long Beach with seasonal ARIMA model. In *2009 International Conference on Energy and Environment Technology*, volume 3, pages 7–10. IEEE.
- Weisberg, S. (2001). Yeo-Johnson power transformations. *Department of Applied Statistics, University of Minnesota*. Retrieved June, 1:2003.
- Westfall, P. H. and Henning, K. S. (2013). *Understanding advanced statistical methods*. CRC Press Boca Raton, FL, USA:.
- White, I. R., Daniel, R., and Royston, P. (2010). Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics & Data Analysis*, 54(10):2267–2275.
- Wickham, H. (2009). *Ggplot2 : Elegant graphics for data analysis*. Springer.
- Wickham, H. (2012). stringr: Make it easier to work with strings. *R Package Version 0.6*, 2:96–7.
- Wickham, H. (2014). tidyr: Easily tidy data with spread () and gather () functions. R package. *Version 0.2.0*. Available at <http://CRAN.R-project.org/package=tidyr> [Verified 7 June 2016].
- Wickham, H. et al. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29.
- Wickham, H. and Francois, R. (2014). dplyr: A grammar of data manipulation (version 0.3.0.2)[software]. *Verfügbar unter http://CRAN.R-project.org/package=dplyr*.

References

- Widiana, D. R., Wang, Y.-F., You, S.-J., Yang, H.-H., Wang, L.-C., Tsai, J.-H., and Chen, H.-M. (2019). Air pollution profiles and health risk assessment of ambient volatile organic compounds above a municipal wastewater treatment plant, Taiwan. *Aerosol Air Qual. Res*, 19:375–382.
- Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.
- Wu, E. M.-Y., Kuo, S.-L., et al. (2012). Air quality time series based GARCH model analyses of air quality information for a total quantity control district. *Aerosol and Air Quality Research*, 12(3):331–343.
- Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798):265–269.
- Xie, C., Huang, C., Zhang, D., and He, W. (2021). Bilstm-i: A deep learning-based long interval gap-filling method for meteorological observation data. *International Journal of Environmental Research and Public Health*, 18(19):10321.
- Xie, J. and Zhu, Y. (2020). Association between ambient temperature and COVID-19 infection in 122 cities from China. *Science of the Total Environment*, 724:138201.
- Xu, Z., Xiong, L., Jin, D., and Tan, J. (2021). Association between short-term exposure to sulfur dioxide and carbon monoxide and ischemic heart disease and non-accidental death in changsha city, china. *Plos one*, 16(5):e0251108.
- Yang, X., Zhang, T., Zhang, X., Chu, C., and Sang, S. (2022). Global burden of lung cancer attributable to ambient fine particulate matter pollution in 204 countries and territories, 1990–2019. *Environmental Research*, 204:112023.
- Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.

References

- Yonar, H., Yonar, A., Tekindal, M. A., and Tekindal, M. (2020). Modeling and forecasting for the number of cases of the COVID-19 pandemic with the curve estimation models, the Box-Jenkins and exponential smoothing methods. *EJMO*, 4(2):160–165.
- Yusof, F., Kane, I. L., and Yusop, Z. (2013). Hybrid of ARIMA-GARCH modelling in rainfall time series. *Jurnal Teknologi*, 63(2).
- Zakaria, N. A. and Noor, N. M. (2018). Imputation methods for filling missing data in urban air pollution data for Malaysia. *Urbanism. Architectura. Constructii*, 9(2):159.
- Zambrano-Monserrate, M. A., García-Albán, F. F., Henk-Vera, K. A., et al. (2016). Bounds testing approach to analyse the existence of an environmental kuznets curve in Ecuador. *International Journal of Energy Economics and Policy*, 6(2):159–166.
- Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50:159–175.
- Zhang, G. P. (2007). A neural network ensemble method with jittered training data for time series forecasting. *Information Sciences*, 177(23):5329–5346.
- Zhang, J. J., Wei, Y., and Fang, Z. (2019). Ozone pollution: a major health hazard worldwide. *Frontiers in Immunology*, 10:2518.
- Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., and Baklanov, A. (2012a). Real-time air quality forecasting, part I: History, techniques, and current status. *Atmospheric Environment*, 60:632–655.
- Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., and Baklanov, A. (2012b). Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects. *Atmospheric Environment*, 60:656–676.
- Zheng, Y., Liu, F., and Hsieh, H.-P. (2013). U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1436–1444.

References

- Zhou, Q., Jiang, H., Wang, J., and Zhou, J. (2014). A hybrid model for $PM_{2.5}$ forecasting based on ensemble empirical mode decomposition and a general regression neural network. *Science of the Total Environment*, 496:264–274.
- Zhou, X., Huang, W., Zhang, N., Hu, W., Du, S., Song, G., and Xie, K. (2015). Probabilistic dynamic causal model for temporal data. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Zhu, L., Hao, Y., Lu, Z.-N., Wu, H., and Ran, Q. (2019). Do economic activities cause air pollution? Evidence from China’s major cities. *Sustainable Cities and Society*, 49:101593.
- Zhu, Y., Xie, J., Huang, F., and Cao, L. (2020). Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China. *Science of the Total Environment*, 727:138704.
- Zou, X., Azam, M., Islam, T., and Zaman, K. (2016). Environment and air pollution like gun and bullet for low-income countries: war for better health and wealth. *Environmental Science and Pollution Research*, 23(4):3641–3657.